ROBUST INFERENCE FOR HETEROGENEOUS TREATMENT EFFECTS WITH APPLICATIONS TO NHANES DATA

by

Ran Mo

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Mathematical Sciences

Indianapolis, Indiana

December 2024

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Honglang Wang, Chair

Department of Mathematical Sciences

Dr. Fang Li

Department of Mathematical Sciences

Dr. Fei Tan

Department of Mathematical Sciences

Dr. Hanxiang Peng

Department of Mathematical Sciences

Approved by:

Dr. Jeffrey X. Watt

To my family.

ACKNOWLEDGMENTS

I would like to begin by expressing my heartfelt gratitude to my advisor, Dr. Honglang Wang. His unwavering guidance, encouragement, and support have been instrumental throughout my Ph.D. research journey. With his insightful advice and mentorship, I have been able to navigate challenges and explore new ideas, ensuring that my research progressed smoothly. The valuable research experience I gained under his supervision has significantly contributed to my growth as a scholar.

I am also deeply thankful to my committee members, Dr. Fei Tan, Dr. Fang Li, and Dr. Hanxiang Peng, for their invaluable feedback and suggestions. Their thoughtful insights have greatly enriched the development and refinement of this dissertation. Their collective expertise has provided me with a well-rounded perspective that I truly appreciate.

In addition, I want to acknowledge the Department of Mathematical Sciences for providing the necessary resources and fostering a supportive environment that greatly facilitated my research. I am particularly grateful to all the professors who have taught me during my academic journey. Their knowledge, passion, and dedication have served as a profound source of inspiration, and the lessons I've learned in their classes have laid a strong foundation for my research and future endeavors. I would also like to extend my gratitude to the administrative staff in the Department of Mathematical Sciences for their steadfast assistance with various matters throughout my Ph.D. journey. Their support in managing both academic and administrative tasks made my experience much smoother, allowing me to focus on my work.

Moreover, I would like to express my appreciation to my friends and colleagues for their advice, encouragement, and camaraderie during this process. Their support has been invaluable, helping me maintain balance and motivation throughout my studies.

Most importantly, I must thank my wife, Dr. Wenxian Zhou, for her endless love, patience, and unwavering support. Her belief in me has been a source of strength and inspiration throughout this journey, and this accomplishment would not have been possible without her by my side.

I also owe a heartfelt thanks to my parents for their lifelong encouragement, love, and sacrifices. Their constant support has been the foundation of my success, and I am forever grateful for their guidance.

This research was supported in part by the National Science Foundation under Grant DMS-2212928 and by the High-Performance Computing resources at IUPUI. I am grateful to the Indiana University Pervasive Technology Institute and Lilly Endowment, Inc., for their financial support.

Lastly, this dissertation utilizes data from the National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics, part of the Centers for Disease Control and Prevention (CDC). I am grateful for the opportunity to work with such valuable data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Estimating the conditional average treatment effect (CATE) using data from the National Health and Nutrition Examination Survey (NHANES) provides valuable insights into the heterogeneous impacts of health interventions across diverse populations, facilitating public health strategies that consider individual differences in health behaviors and conditions. However, estimating CATE with NHANES data face challenges often encountered in observational studies, such as outliers, heavy-tailed error distributions, skewed data, model misspecification, and the curse of dimensionality. To address these challenges, this dissertation presents three consecutive studies that thoroughly explore robust methods for estimating heterogeneous treatment effects.

The first study introduces an outlier-resistant estimation method by incorporating M-estimation, replacing the L_2 loss in the traditional inverse propensity weighting (IPW) method with a robust loss function. To assess the robustness of our approach, we investigate its influence function and breakdown point. Additionally, we derive the asymptotic properties of the proposed estimator, enabling valid inference for the proposed outlier-resistant estimator of CATE.

The method proposed in the first study relies on a symmetric assumption which is commonly required by standard outlier-resistant methods. To remove this assumption while maintaining unbiasedness, the second study employs the adaptive Huber loss, which dynamically adjusts the robustification parameter based on the sample size to achieve optimal tradeoff between bias and robustness. The robustification parameter is explicitly derived from theoretical results, making it unnecessary to rely on time-consuming data-driven methods for its selection. We also derive concentration and Berry-Esseen inequalities to precisely quantify the convergence rates as well as finite sample performance.

In both previous studies, the propensity scores were estimated parametrically, which is sensitive to model misspecification issues. The third study extends the robust estimator from our first project by plugging in a kernel-based nonparametric estimation of the propensity score with sufficient dimension reduction (SDR). Specifically, we adopt a robust minimum average variance estimation (rMAVE) for the central mean space under the potential out-

come framework. Together with higher-order kernels, the resulting CATE estimation gains enhanced efficiency.

In all three studies, the theoretical results are derived, and confidence intervals are constructed for inference based on these findings. The properties of the proposed estimators are verified through extensive simulations. Additionally, applying these methods to NHANES data validates the estimators' ability to handle diverse and contaminated datasets, further demonstrating their effectiveness in real-world scenarios.

1. INTRODUCTION

The National Health and Nutrition Examination Survey (NHANES), conducted by the CDC, gathers comprehensive data on the health and nutritional status of U.S. adults and children through interviews, physical exams, and lab tests, providing essential insights for understanding public health trends and informing policies. Estimating the conditional average treatment effect (CATE) using NHANES data offers a deeper understanding of the heterogeneous impacts of health interventions across diverse populations, enabling more tailored public health strategies. However, the NHANES dataset is prone to outliers, as individuals with certain medical conditions may exhibit values that differ significantly from the general population. Additionally, because medical measurements are often positive, extreme values tend to skew in one direction, leading to potential bias in traditional robust estimation methods. Moreover, model misspecification and high dimensionality also obstacles the CATE estimation in real-world applications.

To address these challenges, this dissertation presents three consecutive studies on robust CATE estimation with applications to NHANES data. These studies aim to improve the reliability of CATE estimation in real-world datasets, where conventional methods often struggle due to deviations from ideal data conditions.

1.1 Outlier Resistant Inference for Conditional Average Treatment Effect

In the first study, a general class of local M-estimation framework (Härdle, 1984) is adopted to enhance the robustness of the well known inverse propensity weighting method for CATE estimation. We evaluate the robustness of our estimator in terms of unbiasedness, influence function, and breakdown point.With derived asymptotic distributions, we are able to perform valid statistical inference, such as constructing confidence intervals.

By employing M-estimators, we relax the requirement for finite conditional moments of potential outcomes as stipulated in Abrevaya, Hsu, and Lieli (2015). Instead, we require finite conditional moments of a robust loss function, a condition typically met by bounded robust loss functions like Hubers loss (Huber, 1964). This modification ensures that our

asymptotic properties hold for observations from heavy-tailed distributions or those affected by outliers, as long as the robust loss function is suitably selected.

Furthermore, we examine the effect of vigorous work activity on urinary albumin levels using data from the 2007-2008 NHANES. The robust CATE estimator is employed to account for outliers in the data. Our results align with existing studies, showing a protective effect of vigorous work activity on albumin levels, particularly for individuals over the age of 60. The narrower confidence intervals from the robust methods provide stronger conclusions regarding the impact of physical activity on albumin excretion.

1.2 Outlier Resistant Inference for Heterogeneous Treatment Effect in The Absence of Symmetry and Light Tail Assumptions

The robust estimator of CATE in our first study is derived under symmetric assumptions. However, traditional M-estimators that assume symmetric errors may not perform well with asymmetric errors. For instance, the conditional median estimated using L_1 loss can differ significantly from the conditional mean in asymmetric models, rendering the outlier-resistant estimator from the first study biased. Notably, the literature lacks approaches for estimating CATE under heavy-tailed and asymmetric potential outcomes. Thus, this paper focuses on developing methods for estimating CATE in these challenging conditions.

Hence in the second study, we adopt the adaptive Huber regression (Sun, W.-X. Zhou, and J. Fan, 2020) for CATE estimation, which dynamically adjusts the robustification parameter based on the sample size. The theoretical properties of our estimator are established through a concentration inequality and a Berry-Esseen type bound. Based on the concentration inequality, the proposed estimator is consistent, requiring only the existence of the second moment for consistency and the third moment for asymptotic normality without assuming symmetric errors while offering enhanced robustness compared to non-robust methods based on squared loss. Since the concentration inequality tends to be too conservative for inference, we propose an asymptotically correct normal-based confidence interval using the Berry-Esseen bound. Under the conditions outlined in the theoretical framework, an explicit form of the robustification parameter is derived, allowing it to be estimated directly without relying on time-consuming data-driven methods.

In this application, we utilize data from the 1999-2006 NHANES to examine the CATE of alcohol consumption on liver function biomarkers, including ALT, AST, GGT, and ALP. Given the heavy-tailed and skewed distribution of the data, our estimator using adaptive Huber loss remains closer to the estimators from non-robust methods compared to other robust methods. At the same time, it provides narrower confidence intervals than non-robust methods. Consequently, findings based on our proposed estimator confirm the positive effects of heavy drinking on all biomarkers, with stronger effects observed in middle-aged individuals compared to younger or older groups, consistent with existing literature. These results underscore the robustness and asymptotically unbiasedness of our method, as well as the importance of accounting for heterogeneity in treatment effects across age groups.

1.3 Nonparametric Outlier Resistant Conditional Average Treatment Effect Estimator with Sufficient Dimension Reduction

In both previous studies, we constructed a robust estimator following the semi-parametric method proposed by Abrevaya, Hsu, and Lieli (2015). where the propensity score, treated as a nuisance parameter, was estimated using parametric methods like logistic regression. However, relying on a parametric model for propensity score estimation can lead to misleading results if the model is misspecified. Alternatively, Abrevaya, Hsu, and Lieli (2015) also proposed a fully nonparametric method, where the propensity score and the CATE estiamtor are both estimated nonparametrically using the Nadaraya-Watson (NW)-estimator(Nadaraya, 1964; Watson, 1964) and higher-order kernels are used to increasing the efficiency. While this approach does not rely on model assumptions, it suffers from the curse of dimensionality, as high-dimensional covariates are often needed to satisfy the ignorability assumption for identifying the CATE estimator.

In the third study, we address these challenges by applying M-estimators to the fully nonparametric CATE estimator proposed by Abrevaya, Hsu, and Lieli (2015) to mitigate issues related to outliers and model misspecification, while incorporating sufficient dimension reduction (SDR) (B. Li, 2018) to minimize the influence of high-dimensional covariates. Specifically, we propose a robust minimum average estimation (rMAVE) approach (Xia et al., 2002) with adjustments for the central mean space under the potential outcomes framework,

and we employ higher-order kernels in the estimation of both the propensity score and CATE to enhance efficiency. Asymptotic theorems are derived, and a comparison of different choices for the central mean space is provided based on these theorems, along with confidence intervals derived from asymptotic normality. The theoretical results, together with the simulation study, reveal that the behavior of the CATE estimator combined with SDR differs from that of the average treatment effect (ATE) (Rosenbaum and Rubin, 1983; Hirano, G. W. Imbens, and Ridder, 2003) estimator with SDR found in the existing literature.

Additionally, we apply our proposed method to analyze data from the 2007-2008 NHANES, focusing on the impact of participation in the National School Lunch Program (NSLP) on body mass index (BMI), conditioned on students' age. With many covariates and a relatively small sample size, our analysis underscores the necessity of our proposed method, as the confidence interval estimated based on our proposed method is obviously narrower than the method without robustness or without dimension reduction. Our findings reveal a significant negative effect on BMI for children under 7, which shifts to a positive effect for those aged 8 and older. These results, which contrast with the non-significant treatment effect reported by Huang and Chan (2017), highlight substantial heterogeneity across age groups.

1.4 Dissertation Structure

The dissertation is structured as follows. In Chapter 2, I present the formulation for robust CATE estimation using M-estimation. Chapter 3 builds upon the first study by introducing adaptive Huber loss to improve performance under skewed error distributions. In Chapter 4, I develop a fully nonparametric robust CATE estimator with sufficient dimension reduction. Each chapter includes relevant literature reviews and methodological details of the proposed approaches.

2. OUTLIER RESISTANT INFERENCE FOR CONDITIONAL AVERAGE TREATMENT EFFECT

The estimation of causal effects based on the CATE is usually vulnerable to outliers. However, to the best of our knowledge, outlier-resistant inference for the CATE has not been investigated in the literature. In this work, we propose an outlier-resistant estimation method for the CATE by incorporating M-estimation into the inverse propensity weighting (IPW) approach. The influence function and breakdown property are investigated to study the robustness of our method. In addition, we derive the asymptotic properties of the proposed estimator for inference purposes. The finite sample performance of the proposed estimator is evaluated via Monte Carlo experiments. The proposed method is compared with the IPW method and the augmented inverse probability weighting (AIPW) method, which do not account for outliers. Finally, the proposed method is applied to the NHANES dataset to estimate the average effects of physical activity on albumin levels conditioned on age.

2.1 Introduction

Treatment effects are widely employed across diverse fields such as healthcare, economics, education, and social sciences to assess the impact of interventions, policies, or treatments on outcomes of interest. Because we do not observe both treated and untreated outcomes for each individual, estimating the individual treatment effect is often unrealistic. A reasonable approach in the literature has been to estimate the average treatment effect (ATE) (Rosenbaum and Rubin, 1983; Hirano, G. W. Imbens, and Ridder, 2003), which provides a single estimate of the treatment's impact on an entire population, thereby offering valuable insights into the overall effectiveness of the intervention.

With the growing interest in estimating and analyzing heterogeneous treatment effects in both experimental and observational studies, researchers are increasingly focusing on the conditional average treatment effect (CATE). Traditional methods include regression adjustment (Rosenbaum and Rubin, 1983), propensity score matching (Rosenbaum and Rubin, 1983), inverse probability weighting (IPW) (Hirano, G. W. Imbens, and Ridder, 2003),

and augmented inverse probability weighting (AIPW) (Glynn and Quinn, 2010; Kurz, 2022). Beyond these traditional approaches, researchers have adopted a variety of methodologies, including meta-learners such as the X-learner (Künzel et al., 2019), T-learner (Athey and G. Imbens, 2016; L. Yao et al., 2018; Powers et al., 2018), R-learner (Nie and Wager, 2021), and S-learner (Dorie et al., 2019; Foster, Taylor, and Ruberg, 2011). These methods decompose CATE estimation into manageable sub-regression problems, allowing for the integration of non-causal approaches. Additionally, there are algorithms specifically designed for CATE estimation, such as causal forests (Athey, Tibshirani, and Wager, 2019) and Bayesian additive regression trees (BART) (Hill, 2011; Chipman, George, and McCulloch, 2010).

In practical applications, researchers may only be interested in the CATE of subpopulations conditioned on a subset of all covariates. The definition of CATE has been extended by Abrevaya, Hsu, and Lieli, 2015 to address a more technically challenging situation where the conditioning covariates X_1 are continuous and form a strict subset of X. Since the unconfoundedness assumption does not generally hold when conditioning solely on X_1, it is inadequate to simply apply, for example, the CATE estimator from Lee and Whang (2009) using X_1 in place of X. Instead, one must estimate CATE as a function of X and subsequently average out the unwanted components of X not included in X_1. However, this distribution is typically unknown and must be estimated. Notable CATE estimators include propensity-score-based methods, such as the IPW estimator (Abrevaya, Hsu, and Lieli, 2015) and the doubly robust (DR) estimator (Lee, Okui, and Whang, 2017; Q. Fan et al., 2022). These methods often condition on a subset of all covariates, resulting in what is known as the reduced dimensional conditional average treatment effect (Q. Fan et al., 2022).

However, estimating treatment effects in real-world scenarios often involves dealing with data contaminated by outliersextreme data points that deviate significantly from the majority of observations. When outliers are present in the response variable, estimators based on the sample mean tend to be heavily influenced by these outliers. While robust techniques for estimating ATE have been studied (Firpo, 2007; Z. Zhang et al., 2012; Harada and Fujisawa, 2021), there is a gap in the literature regarding outlier-resistant estimators for the CATE. Existing methods such as those proposed by Abrevaya, Hsu, and Lieli (2015), Lee, Okui,

and Whang (2017), and Q. Fan et al., 2022 exhibit limited outlier resistance due to their reliance on the L_2 losses.

To address the challenge of outliers in the response variable, traditional methods in robust statistics, such as M-estimators, have frequently been used. Originally proposed by Huber, 1964 as an outlier-resistant location estimator and later introduced with kernel regression by Härdle, 1984, M-estimators seek to find a parameter estimate that minimizes the sum of robust loss functions of the data, rather than relying solely on the mean or median. The robust function is designed to reduce susceptibility to outliers while maintaining efficiency.

Building on these traditional robust methods, this paper extends the second step of the semiparametric method in Abrevaya, Hsu, and Lieli, 2015 from a NW estimator to a general class of local M-estimators (Härdle, 1984). We discuss the robustness of our estimator from the perspectives of unbiasedness, influence function (IF), and breakdown point. Asymptotic properties are derived, and confidence intervals are constructed based on the mean and variance estimators from asymptotic distributions.

The application of M-estimators allows us to relax the condition of finite conditional moments of the potential outcomes as required in Abrevaya, Hsu, and Lieli, 2015. Instead, we require finite conditional moments of a robust loss function of the potential outcomes. This modified condition is typically satisfied for bounded robust loss functions, such as the Huber loss function (Huber, 1964). Theoretically, this guarantees that asymptotic properties hold for observations from heavy-tailed distributions or those contaminated by outliers, provided that the robust loss function is appropriately chosen.

To verify our theoretical results, we conduct Monte Carlo simulations to study and illustrate our method. The proposed estimators perform well in terms of bias, variance, and MSE when the generated data is contaminated. The validity of the asymptotic distribution and variance is confirmed through the constructed confidence intervals.

Furthermore, we examine the effect of vigorous work activity on urinary albumin levels using data from the 20072008 National Health and Nutrition Examination Survey (NHANES). The robust CATE estimator is employed to account for outliers in the data. Our results align with existing studies, showing a protective effect of vigorous work activity on albumin levels, particularly for individuals over the age of 60. The narrower confidence intervals from

the robust methods provide stronger conclusions regarding the impact of physical activity on albumin excretion.

The rest of the article is organized as follows. Section 2 introduces the outlier-resistant CATE estimator and discusses its identification and estimation. The algorithm for our proposed estimator is also presented in this section. In Section 3, we demonstrate the outlier resistance of our estimator by examining its behavior from the perspectives of the influence function and the breakdown point. Section 4 develops the asymptotic properties of the proposed estimators. Section 5 presents the simulation results, and Section 6 is devoted to the empirical exercise. Section 7 concludes the paper.

2.2 Methodology

2.2.1 Problem Setup

Under the potential outcome framework (Rubin, 1974; Neyman, 1923), let $Y^{(1)}$ and $Y^{(0)}$ denote the potential outcomes with and without treatment, respectively. For each individual, the observed outcome Y is defined by

$$Y = TY^{(1)} + (1-T)Y^{(0)},$$

where $T \in \{0,1\}$ is an indicator variable of the treatment state, with $T = 1$ if an individual receives treatment and $T = 0$ otherwise. Let X be a p-dimensional vector of covariates with $p \geq 2$. Given Y, T, and X, the observations are given by $\{(T_{\mathrm{i}}, Y_{\mathrm{i}}, X_{\mathrm{i}}), \mathrm{i} = 1, 2, \ldots, n\}$ from the independent and identically distributed joint distribution of the vector (T, X, Y).

The reduced dimensional conditional average treatment effect is defined as

$$\tau(x_1) = E[Y^{(1)} - Y^{(0)} \mid X_1 = x_1] = \mu_1(x_1) - \mu_0(x_1), \tag{2.1}$$

where X_1 is a fixed d-dimensional subvector of the covariates X, with $d \leq p$, and the conditional means of the potential outcomes are defined as

$$\mu_1(x_1) = E[Y^{(1)} \mid X_1 = x_1], \quad \mu_0(x_1) = E[Y^{(0)} \mid X_1 = x_1].$$

2.2.2 Identify CATE

Since we cannot estimate CATE directly from (2.1) because the potential outcomes in the treatment and control groups cannot be observed simultaneously, we need to identify our parameter of interest with the observed outcome Y instead of the potential outcomes.

Following Abrevaya, Hsu, and Lieli (2015), we consider the situation where the unconfoundedness assumption does not generally hold when conditioning on the low-dimensional covariates X_1. The identification of the estimand $\mu_1(x_1)$ and $\mu_0(x_1)$ is based on the common assumptions (Rosenbaum and Rubin, 1983):

Assumption 2.2.1.

1. (Unconfoundedness) $(Y^{(1)}, Y^{(0)}) \perp T \mid X$.

2. (Positivity) Let $\boldsymbol{\pi}(x) = P(T = 1 \mid X = x)$ *be the propensity score, where there exists* $C > 0$ *such that* $P(C \leq \boldsymbol{\pi}(X) \leq 1 - C) = 1$.

Based on Assumption 2.2.1, the IPW method proposed by Abrevaya, Hsu, and Lieli (2015) is based on the following one-model estimating equation:

$$0 = E\left[\frac{TY}{\boldsymbol{\pi}(X)} - \frac{(1-T)Y}{1-\boldsymbol{\pi}(X)} - \tau(x_1) \mid X_1 = x_1\right].$$

And the estimator $\hat{\tau}(x_1)$ is found using its empirical form:

$$0 = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}\left[\frac{T_{\mathrm{i}}Y_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})} - \frac{(1-T_{\mathrm{i}})Y_{\mathrm{i}}}{1-\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})} - \hat{\tau}(x_1)\right] K\left(\frac{X_{\mathrm{i}1} - x_1}{h}\right), \tag{2.2}$$

where $K(\cdot)$ is a kernel function with bandwidth h and $\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})$ is the estimator of propensity score.

With outliers for the response, a direct approach to incorporate (2.2) with the M-estimator might involve truncating the residual $\frac{T_{\mathrm{i}}Y_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})} - \frac{(1-D_{\mathrm{i}})Y_{\mathrm{i}}}{1-\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})} - \hat{\tau}(x_1)$ with a robust function such as the Huber loss function.

However, since the contaminated distribution is defined in the counterfactual observations, not directly in $\frac{T_{\mathrm{i}}Y_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})} - \frac{(1-T_{\mathrm{i}})Y_{\mathrm{i}}}{1-\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}$, simply truncating the residuals in (2.2) to remove terms with large deviations or reduce their influence does not seem reasonable. To derive an outlier-

resistant method based on the idea of M-estimator, we need to truncate the direct difference of potential outcomes and their mean functions, $(Y_{\mathrm{i}} - \mu_1(x_1))$ when $T = 1$ or $(Y_{\mathrm{i}} - \mu_0(x_1))$ when $T = 0$, in the estimating equation.

$$\tau(x_1) = \mu_1(x_1) - \mu_0(x_1).$$

where $\mu_1(x_1) = E[Y^{(1)} \mid X_1 = x_1]$ and $\mu_0(x_1) = E[Y^{(0)} \mid X_1 = x_1]$ denote the conditional means of each counterfactual outcome, and the corresponding estimator $\hat{\tau}(x_1)$ can be calculated by:

$$\hat{\tau}(x_1) = \hat{\mu}_1(x_1) - \hat{\mu}_0(x_1).$$

In this study, our demonstration will focus solely on μ_1, as μ_0 can be addressed in a similar manner.

Here we provide the identification of $\mu_1(x_1)$ based on Assumption 2.2.1. Under the unconfoundedness and positivity assumptions, it is easy to see

$$0 = E\left[\frac{T}{\boldsymbol{\pi}(X)}\left(Y - \mu_1(x_1)\right) \mid X_1 = x_1\right].$$

With this, the solution $\hat{\mu}_1(x_1)$ to its empirical form can be regarded as a local constant estimator with additional weight $\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}$:

$$0 = \sum_{\mathrm{i}=1}^{n} \frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}(Y_{\mathrm{i}} - \hat{\mu}_1(x_1))K\left(\frac{X_{\mathrm{i}1} - x_1}{h}\right).$$

2.2.3 Outlier-Resistant Estimation Algorithm

To derive a robust estimator based on the estimating equation, we replace the term $(Y - \mu)$ with an anti-symmetric function $\psi(Y - \mu)$ and modify the estimating equation as follows:

$$0 = E\left[\frac{T}{\boldsymbol{\pi}(X)}\psi(Y - \mu) \mid X_1 = x_1\right]. \tag{2.3}$$

The symmetry assumption is widely adopted as a prerequisite for employing classical robust methods (Rousseeuw et al., 1986), which tend to ignore or put less weight on extreme

observations on both sides of the mean, such as the sample median or Huber regression, as estimators for the population mean.

Assumption 2.2.2. *The robust function $\psi(\cdot)$ is antisymmetric and the conditional density $f(y^{(1)} \mid x_1)$ is symmetric with respect to $\mu_1(x_1)$.*

Theorem 2.2.1. Under Assumptions 2.2.1 and 2.2.2, the solution to the equation (2.3) is $\mu_1(x_1)$.

Since $\mu_1(x_1)$ is a solution to the estimating equation (2.3), we propose to estimate $\hat{\mu}_1(x_1)$ by solving the empirical form of the estimating equation:

$$0 = \sum_{\mathrm{i}=1}^{n} \frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})} \psi\left(Y_{\mathrm{i}} - \hat{\mu}_1(x_1)\right) K\left(\frac{X_{\mathrm{i}1} - x_1}{h}\right). \tag{2.4}$$

The solution to equation (2.4) can be viewed as a weighted M-estimator with weight $W_{\mathrm{i}} = \frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} K(\frac{X_{\mathrm{i}1} - x_1}{h})$. Therefore, we can solve (2.4) using the iteratively reweighted least squares (IRLS) algorithm (Holland and Welsch, 1977) with the following iterative procedure.

From $\hat{\mu}_1^{(t)}(x_1)$ obtained from the t-th step we solve the $\hat{\mu}_1^{(t+1)}(x_1)$ from the equation

$$0 = \sum_{\mathrm{i}=1}^{n} \frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} K\left(\frac{X_{\mathrm{i}1} - x_1}{h}\right) W(Y_{\mathrm{i}} - \hat{\mu}_1^{(t)}(x_1))(Y_{\mathrm{i}} - \mu_1^{(t+1)}(x_1)), t = 0, 1, 2, ...,$$

where

$$W(x) = \begin{cases} \frac{\psi(x)}{x} & \text{if } x = 0, \\ \psi'(x) & \text{if } x = 0. \end{cases}$$

Furthermore, the propensity score is typically unknown in observational studies. As suggested by the semiparametric method of Abrevaya, Hsu, and Lieli (2015), we assume the propensity score is correctly specified by the generalized linear model and estimated using a parametric method, such as logistic regression.

2.3 Robustness

In this subsection, we demonstrate the outlier-resistance of our proposed estimator by examining its behavior from the perspectives of the influence function and the breakdown point.

2.3.1 Influence Function

Introduced by Hampel, 1974, the influence function (IF) of an estimator provides an approximation of how the estimator behaves when the sample contains a small fraction ϵ of identical outliers. Follow Harada and Fujisawa, 2021, we demonstrate the outlier-resistant property of our proposed estimator $\hat{\mu}_1(x_1)$ from the viewpoint of the influence function. Let $\theta[F]$ denote the parameter of interest θ based on a sample from a distribution F, then the influence function of an estimator θ at distribution F is defined as:

$$IF_{\theta(x_0,F)} = \lim_{\varepsilon \to 0} \frac{\theta[(1-\varepsilon)F + \varepsilon\delta_{x_0}] - \theta[F]}{\varepsilon} = \frac{\partial}{\partial \epsilon} \left\{\theta[(1-\epsilon)F + \epsilon\delta_{x_0}]\right\} \mid_{\epsilon=0},$$

where δ_x denotes the Dirac delta function.

Back to our problem, we denote our parameter of interest under contaminated distribution by $\mu_\epsilon(x_1) = \mu_1[\tilde{g}](x_1)$, where $\tilde{g}$ is the contaminated density function of $Y \mid X = x, T = t$ given by

$$\tilde{g}(y \mid x, t) = (1-\epsilon)g(y \mid x, t) + \epsilon\delta_{y_0},$$

where $g(y \mid x, t)$ denotes the uncontaminated density function of $Y \mid X = x, T = t$. Under the contaminated distribution $\tilde{g}(y \mid x, t)$, denote $\mu_\epsilon(x_1)$ as the solution to our estimating equation:

$$0 = E_{\tilde{g}}\left[\frac{T}{\boldsymbol{\pi}(X)}\psi(Y - \mu_\epsilon(x_1)) \mid X_1 = x_1\right].$$

To find the influence function, we first rewrite the equation in terms of potential outcomes:

$$0 = E_{\tilde{g}}\left[\psi(Y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1\right].$$

Next, we calculate the partial derivative with respect to ϵ:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \epsilon} E_{\tilde{g}} \left[\psi(Y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1 \right] \\
0 &= \frac{\partial}{\partial \epsilon} \iiint \psi(y^{(1)} - \mu_\epsilon(x_1)) \left[(1-\epsilon) g(y^{(1)} \mid t, x) + \epsilon \delta_{y_0} \right] g(t \mid x) g(x \mid x_1) \, dy \, dt \, dx_1^c \\
0 &= -E_g \left[\psi(Y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1 \right] - (1-\epsilon) \frac{\partial \mu_\epsilon(x_1)}{\partial \epsilon} E_g \left[\psi'(Y^{(1)} - \mu_\epsilon(x_1)) \right] \\
&\quad + \iiint \psi(y^{(1)} - \mu_\epsilon(x_1)) \delta_{y_0} g(t \mid x) g(x \mid x_1) \, dy \, dt \, dx_1^c \\
&\quad - \epsilon \frac{\partial \hat{\mu}_\epsilon}{\partial \epsilon} \iiint \psi'(y^{(1)} - \mu_1(x_1)) \delta_{y_0} g(y^{(1)} \mid t, x) g(t \mid x) g(x \mid x_1) \, dy \, dt \, dx_1^c.
\end{aligned}$$

As $\epsilon \to 0$, the first and last terms vanish, and $\frac{\partial \mu_\epsilon(x_1)}{\partial \epsilon}$ corresponds to the influence function. Taking the limit $\epsilon \to 0$, we obtain:

$$0 = -IF_{(y_0, \tilde{g})} E_g \left[\psi'(y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1 \right] + \psi(y_0 - \mu_\epsilon(x_1)).$$

Therefore,

$$IF_{(y_0, \tilde{g})} = \left[E_g \left[\psi'(y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1 \right] \right]^{-1} \psi(y_0 - \mu_\epsilon(x_1)).$$

When $y_0 \to \infty$, the value of the influence function is bounded if $\psi(\cdot)$ is bounded or approaches 0 for certain re-descending loss functions. These properties are advantageous for outlier resistance.

2.3.2 Breakdown Point

In this section, we're going to discuss the breakdown point of a location estimator with an estimating equation of the form:

$$0 = \sum_{\mathrm{i}=1}^{n} \psi(y_\mathrm{i} - \mu) W_\mathrm{i},$$

where the weights are defined as $W_\mathrm{i} = \frac{T_\mathrm{i}}{\pi(X_\mathrm{i})} K(\frac{X_{\mathrm{i}1} - x_1}{h})$.

To imitating the contamination of a sample, there are two approaches: replacement contamination and addition contamination (Donoho and Huber, 1983). In this subsection, we

adopt the replacement contamination approach to define the finite sample breakdown point. Specifically, we consider a sample of size n and randomly replace m of the observations with corrupted ones. For this subsection, let S_{Y_n} denote the set of the outcomes in noncontaminated observations and S_{Y_c} the set of the outcomes of contaminated observations, the set of all responses is denoted by S_Y. Then the estimating equation can thus be rewritten as:

$$0 = \sum_{y_{\mathrm{i}} \in S_{Y_n}} \psi(y_{\mathrm{i}} - \mu) W_{\mathrm{i}} + \sum_{y_{\mathrm{i}} \in S_{Y_c}} \psi(y_{\mathrm{i}} - \mu) W_{\mathrm{i}}.$$

Since the weights are notably related to T_{i} and X_{i}, a fixed global breakdown point cannot be determined. Following this existing study Giloni and Simonoff (2005), we define the breakdown point of our estimator in a local and conditional manner.

Then we can define the conditional breakdown point as:

$$\varepsilon(x_1 \mid X, T) = \min_{0 \le m^* < n/2} \left\{ \frac{m^*}{n} \mid \sup_{y_{\mathrm{i}} \in S_{Y_c}} \mid \hat{\mu}(S_{Y_n} \cup S_{Y_c}) - \hat{\mu}(Y) \mid = \infty \text{ with fixed } X, T, x_1 \right\}.$$

To study the breakdown property, G. Li and Jian Zhang (1998) categorized the general class of loss functions ρ into three different types: $B\rho$, $U\rho$, and $C\rho$. Since our ψ function can be regarded as the derivative of a differentiable ρ function, we can follow the categories in G. Li and Jian Zhang (1998) and study the breakdown property of our estimator $\hat{\tau}(x_1)$ in the following conditions.

Condition 1: $(B\rho)$ $\rho(x)$ attains its minimum -1 at $x = 0$; ρ is non-increasing for $x < 0$ and non-decreasing for $x > 0$. Furthermore, $\rho(x) \to 0$ as $\mid x \mid \to \infty$.

Condition 2: $(U\rho)$ $\rho(x)$ attains its minimum 0 at $x = 0$; $\rho(x)$ is symmetric about 0; $\rho(x)$ is non-decreasing for $x > 0$ and $\lim_{|x| \to \infty} \rho(x) = \infty$; $\psi = \rho'$ is continuous in $\mathbb{R}^1$, and there exists $x_0 \ge 0$ such that ψ is non-decreasing in $(0, x_0]$ and non-increasing in (x_0, ∞).

Condition 3: $(C\rho)$ $\rho(x)$ is convex, $\psi(x) := \rho'(x)$ exists everywhere and $\psi(-\infty) < 0 < \psi(+\infty)$.

The three conditions outlined above include many commonly used loss functions $\rho(x)$, such as the L_2 loss, the L_1 loss, Huber loss (Huber, 1973), and Tukey's biweighted loss (Beaton and Tukey, 1974).

Under condition 1, the derivative of the loss function approaches 0 as x goes to infinity, and the loss function remains bounded. All bounded redescending M-estimators, such as the M-estimator based on Tukey's biweighted loss, can be rescaled to satisfy this condition. Condition 2 encompasses symmetric unbounded redescending M-estimators. Specifically, it includes M-estimators where the derivative of the loss function is bounded and bounded away from zero, while the loss function itself is unbounded. This class of M-estimators includes those based on the Huber loss, the L_1 loss, and other loss functions that tend to infinity at a slower rate than the Huber and L_1 losses. Finally, condition 3 includes the Huber loss, the L_1 loss, and other loss functions that diverge more rapidly than the Huber and L_1 losses.

Based on these conditions, we have the following Theorem:

Theorem 2.3.1. *(1) Under condition 1 ($B\rho$), define* $A = -\Sigma_{y_i \in S_{Y_n}} \rho(y_i - \mu) W_i \Big/ \Sigma_{y_i \in S_{Y_n}} W_i$, *Then the estimator will be unbounded if:*

$$\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} < \frac{1-A}{A},$$

and will be finite when:

$$\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} > \frac{1-A}{A}.$$

(2) Under the second condition($U\rho$), the estimator will be bounded if

$$\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} > 1,$$

and will be unbounded if

$$\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} < 1.$$

(3) Under condition 3 ($C\rho$), assume that $\psi(-\infty) = k_1$ and $\psi(+\infty) = k_2$. Then the estimator μ will be bounded if

$$\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} \geq \max\left\{\frac{k_1}{k_2}, \frac{k_2}{k_1}\right\},$$

and will be unbounded if

$$\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} < \max\left\{\frac{k_1}{k_2}, \frac{k_2}{k_1}\right\}.$$

The theorem indicates that the breakdown properties of the estimator are closely related to the ratio of the summation of the weights, $\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i}$. For each case, the estimator will break down when this ratio falls below a certain threshold or be finite when this ratio falls over a certain threshold. This shows that the breakdown point is closely related to the minimum number m of the greatest weights chosen into the S_{Y_c} group that satisfies the breakdown condition corresponding to its loss function, that is:

$$\varepsilon(x_1 \mid X, T) = \min_{0 \le m^* < \frac{n}{2}} \left\{ \frac{m^*}{n} \mid \sup_{S_{Y_c}} \frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} < C \right\},$$

where $W_i = \frac{T_i}{\hat{\pi}(X_i)} K(\frac{X_{i1} - x_1}{h})$ and C is a constant depend on theorem 2.3.1.

In real-world applications, even though contaminated observations are randomly chosen, it is reasonable to consider the worst-case scenario where the observations with the highest weights are contaminated. When some of the W_i's are significantly larger than others, a few contaminated observations might be sufficient to cause the estimator to break down. Given that the weights are expressed as $W_i = \frac{T_i}{\hat{\pi}(X_i)} K\left(\frac{X_{i1} - x_1}{h}\right)$, to mitigate breakdown issues, we can consider either increasing the bandwidth h in the kernel function or truncating observations with extreme estimated propensity scores.

2.4 Asymptotic Properties

In this section, we develop the asymptotic theory for our proposed estimator. We begin by establishing the asymptotic properties of $\hat{\mu}_1(x_1)$ and then extend our results to derive the asymptotic properties of $\hat{\tau}(x_1)$. Additionally, we provide the necessary variance estimator for constructing a point-wise confidence interval.

Our asymptotic Theorems are established under the following underlying assumptions:

Assumption 2.4.1. *(Estimated propensity score). The propensity score is correctly specified and estimated by parametric method with* $sup_{x \in \mathcal{X}} \mid \hat{\boldsymbol{\pi}}(X) - \boldsymbol{\pi}(X) \mid = O_p(1/\sqrt{n})$.

For simplicity, we assume that the propensity score converges at a rate fast enough for its estimation error to be dominated by the error of the parameter of interest as $n \to \infty$. Assumption 2.4.1 typically holds for standard parametric estimation methods, such as logistic regression or maximum likelihood with a probit model, under reasonably mild regularity conditions.

Assumption 2.4.2. *(Robust function). In addition to (A2), $\psi(u)$ is monotone, bounded and having two continuous bounded derivatives with $\psi'(0) > 0$.*

Assumption 2.4.3. *(Distribution of Y and X_1). The conditional probability density function $f(Y^{(1)} \mid X_1)$ have bounded partial derivative on $x_1 \in \mathcal{X}$ a Cartesian product of compact intervals of the real line. The density of X_1, is assumed to be positive on $\mathcal{X}$.*

Assumption 2.4.4. *(Conditional moments and smoothness).* $\sup_{x_1 \in \mathcal{X}} E[\psi(Y^{(\mathrm{j})} - \mu_{\mathrm{j}})^2 \mid X_1 = x_1] < \infty$ *for* $\mathrm{j} = 0, 1$. *The functions* $m_{\mathrm{j}}(x_1) = E[\psi(Y^{(\mathrm{j})} - \mu_{\mathrm{j}}) \mid X_1 = x_1], \mathrm{j} = 0, 1$ *and $g(x_1)$ the density function of X_1 is $s \geq 2$ times continuously differentiable.*

Assumption 2.4.5. *(Kernel). $K(u)$ is a kernel of order s symmetric around 0 with bouned support $[-A, A]$ and $\int u^2 K(u) du < \infty$.*

Assumption 2.4.6. *(Bandwidths). The bandwidths h satisfy the conditions: $nh^{2s+d} \to 0$ and $nh^d \to \infty$ as $n \to \infty$.*

For convenience, we apply the notation and results from Härdle (1984). Let $\delta_n(\cdot)$ denote the delta function sequence (DFS), which satisfies

(D1) $\int \mid \delta_n(u) \mid du < \infty$, for all n,

(D2) $\int \delta_n(u) du = 1$,

(D3) $\delta_n(u) \to 0$, uniformly on $\mid u \mid > \eta, \eta > 0$ as $n \to \infty$,

(D4) $\int_{|u|>\eta} \delta_n(u) du \to 0$, for each $\eta > 0$ as $n \to \infty$.

Our choice of kernel $\delta(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$ can be regarded as a DFS of kernel type.

For simplicity, we will demonstrate the case when $d = 1$ and $s = 2$, as the proof can be easily extended to higher orders and dimensions.

2.4.1 Consistency

Based on the definition of DFS, the estimating equation can be written as:

$$\frac{1}{n}\sum_{\mathrm{i}=1}^{n}\psi(Y_{\mathrm{i}}-\theta)\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}\delta_n(x_1-X_{\mathrm{i}1})=0,$$

We define the left-hand side of the estimating equation as:

$$h(x_1,\mu)=\frac{1}{n}\sum_{\mathrm{i}=1}^{n}\psi(Y_{\mathrm{i}}-\mu)\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}\delta_n(x_1-X_{\mathrm{i}1}),$$

and define

$$H(x_1,\mu)=E\left[\psi(Y_{\mathrm{i}}-\mu)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\mid X_{\mathrm{i}1}=x_1\right]g(x_1),$$

We first demonstrate the consistency of the estimating equation.

Lemma 2.4.1. *Given assumptions 2.2.1 to 2.4.6, we have* $h(x_1,\mu)\xrightarrow{P}H(x_1,\mu)$.

Building on Lemma 2.4.1, we can establish the consistency of $\hat{\mu}_1(x_1)$ when the loss function ψ is monotone.

Theorem 2.4.1. *Suppose* $\psi(\cdot)$ *is continuous, strictly increasing, and there exists a constant* c *such that* $H(x_1,\mu)>0$ *for* $s>c$ *and* $H(x_1,\mu)<0$ *for* $s<c$. *Then, under the assumptions in Lemma 2.4.1,* $\hat{\mu}_1(x_1)\xrightarrow{P}\mu_1(x_1)$.

Remark 2.4.1. *For continuous but non-monotone* ψ *functions, such as Tukey's biweight loss, we define the estimator based on the estimator* $\hat{\mu}_1(x_1)$ *from monotone* ψ *functions:*

$$\mid\hat{\mu}_1(x_1)-\tilde{\mu}_1(x_1)\mid=\inf_{\mu}\{\mid\mu-\hat{\mu}_1(x_1)\mid:h(x_1,\mu)=0\}.$$

That is the solution to the estimating equation with the non-monotone ψ *function that is closest to* $\hat{\mu}_1(x_1)$ *our estimator in Theorem 2.4.1 with monotone* ψ *function.*

By selecting the solution nearest to $\hat{\mu}_1(x_1)$, *we obtain the consistency of* $\tilde{\mu}_1(x_1)$

Corollary 2.4.1. *Suppose that the assumptions of Theorem 2.4.1 hold, then* $\tilde{\mu}_1(x_1)\xrightarrow{P}\mu_1(x_1)$.

2.4.2 Asymptotic Normality

To formulate the result of the asymptotic normality, let us define:

$$Z_n(x_1) = \frac{C_1(x_1)\left[\hat{\mu}_1(x_1) - \mu_1(x_1) - \frac{B_n(x_1)}{C_1(x_1)g(x_1)}\right]}{\left[\frac{\alpha_n(2)}{n}\sigma^2(x_1)g^{-1}(x_1)\right]^{1/2}},$$

where

$\sigma^2(x_1) = E\left[\psi^2(Y - \mu_1(x_1))\frac{T}{\pi(X)^2} \mid X_1 = x_1\right]$,

$C_1(x_1) = E(\psi'(y - \mu_1(x_1)) \mid X_1 = x_1)$,

$B_n(x_1) = E(h(x_1, \mu_1(x_1)) \mid X_1 = x_1)$,

$\alpha_n(2) = \int \delta_n(u)^2 du$.

Then fllow on the definitions, we have

Theorem 2.4.2. *Suppose the conditions in Theorem 2.4.1 or Corollary 2.4.1 hold, we have* $Z_n \xrightarrow{D} N(0,1)$ *if further satisfies the properties:*

(1) $\gamma_n = \int \mid \delta(u) \mid^{2+\eta} du < \infty$ *for some* $\eta > 0$,

(2) $\gamma_n = o(n^{\eta/2}\alpha_n(2)^{1+\eta/2})$ *as* $n \to \infty$.

Theorem 2.4.1 and Theorem 2.4.2 show that $\hat{\mu}_1(x_1)$ converges in probability and has a normal limiting distribution. However, Theorem 2.4.2 does not directly assist in constructing a confidence interval, as the normal limiting distribution may not have a mean of zero. To eliminate the bias term from the limiting distribution, we must ensure that

$$\frac{B_n(x_1)}{\left[\frac{\alpha_n(2)}{n}\sigma^2(x_1)g(x_1)\right]^{1/2}} \to 0,$$

as $n \to \infty$, which means we need $B_n(x_1) = o((\alpha_n/n)^{-1/2})$.

Since K is a kernel of order s, we have $\int u^t K(u)du = 0$ for $t \in \{1, 2, ..., s\}$ then by Taylor expansion, we have:

$$\begin{aligned} B_n(x_1) &= E\left\{(\psi^{(1)} - E[\psi^{(1)} \mid X])\frac{T}{\pi(X)} + E[\psi^{(1)} \mid X] \mid X_1 = x_1\right\} g(x_1) + O(h^s) \\ &= E\left[\psi(Y^{(1)} - \mu_1(x_1)) \mid X_1 = x_1\right] g(x_1) + O(h^s). \end{aligned}$$

The first term can be interpreted as the bias due to the application of function ψ which equals 0 by the unbiasedness of the estimating equation as we proved in Theorem 2.2.1, we have

$$B_n(x_1) \quad = \quad O(h^s).$$

We only need $O(h^s) = o\left(1/\sqrt{nh}\right)$, which is equivalent to $nh^{2s+1} \to 0$. When x_1 is l dimensional, the condition becomes $nh^{2s+l} \to 0$, which coincides with assumption 2.4.6. Under these conditions, the bias term can be dropped, and a confidence interval can be provided accordingly.

Theorem 2.4.3. *When the conditions of Theorem 2.4.2 hold, we have*

$$Z_n^*(x_1) = \frac{C_1(x_1)\left(\hat{\mu}_1(x_1) - \mu_1(x_1)\right)}{\left[\frac{\alpha_n}{n}\sigma^2(x_1)g^{-1}(x_1)\right]^{1/2}} \xrightarrow{D} N(0,1).$$

2.4.3 Asymptotic Properties of

Based on Theorem 2.4.1, we can similarly extend the asymptotic properties to $\hat{\mu}_0(x_1)$, assuming that the same kernel and bandwidth are used for both $\mu_1(x_1)$ and $\mu_0(x_1)$. By subtracting the two separate estimators, we achieve the consistency result: $\hat{\tau}(x_1) \xrightarrow{P} \tau(x_1)$.

And the normality:

$$\sqrt{\frac{ng(x_1)}{\alpha_n(2)\sigma^2(x_1)}}(\hat{\tau}(x_1) - \tau(x_1)) \xrightarrow{D} N(0,1),$$

where

$$\sigma^2(x_1) = E\left[\left(\frac{H_n^{*1}(x_1)}{D^1(x_1)} - \frac{H_n^{*0}(x_1)}{D^0(x_1)}\right)^2 \mid X_1 = x_1\right].$$

2.4.4 The Variance Estimator and Confidence Interval

Based on the asymptotic distribution, we can get the standard error by

$$SE(\hat{\tau}(x_1)) = \sqrt{\frac{\alpha_n(2)\hat{\sigma}^2(x_1)}{n\hat{g}(x_1)}},$$

where

$$\hat{\sigma}^2(x_1) = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\left(\frac{\psi(Y_{\mathrm{i}}-\hat{\mu}_1(x_1))\frac{T_{\mathrm{i}}}{\hat{\pi}(X_{\mathrm{i}})}}{\hat{D}^1(x_1)} - \frac{\psi(Y_{\mathrm{i}}-\hat{\mu}_0(x_1))\frac{1-T_{\mathrm{i}}}{1-\hat{\pi}(X_{\mathrm{i}})}}{\hat{D}^0(x_1)}\right)^2 \Bigg/ \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right),$$

with

$$\hat{D}^1(x_1) = \frac{1}{nh_2}\sum_{\mathrm{i}=1}^{n} K\left(\frac{X_{\mathrm{i}1}-x_1}{h_2}\right)\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}\psi'(Y_{\mathrm{i}}-\hat{\mu}_1(x_1)) \Bigg/ \frac{1}{nh_2}\sum_{\mathrm{i}=1}^{n} K\left(\frac{X_{\mathrm{i}1}-x_1}{h_2}\right),$$

$$\hat{D}^0(x_1) = \frac{1}{nh_3}\sum_{\mathrm{i}=1}^{n} K\left(\frac{X_{\mathrm{i}1}-x_1}{h_3}\right)\frac{1-T_{\mathrm{i}}}{1-\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}\psi'(Y_{\mathrm{i}}-\hat{\mu}_0(x_1)) \Bigg/ \frac{1}{nh_3}\sum_{\mathrm{i}=1}^{n} K\left(\frac{X_{\mathrm{i}1}-x_1}{h_3}\right).$$

The bandwidths h_1, h_2, and h_3 are specific to the kernel regressions in the estimation of variance of robust CATE estimator and must be chosen independently in practice to balance bias and variance trade-offs effectively.

Together with our proposed point estimator, this allows us to construct a normal confidence interval.

2.5 Monte Carlo Simulations

In this section, we evaluate the finite sample performance of the proposed estimators through Monte Carlo simulations. We first assess the accuracy of the asymptotic approximations by estimating the CATE using fixed bandwidths. Next, we estimate the CATE again using bandwidths selected via cross-validation. We compare these results with non-robust

estimators, specifically the existing non-robust IPW-based estimators (Abrevaya, Hsu, and Lieli, 2015).

2.5.1 Data Generating Process

The data for this study is generated using a simple causal framework. The covariates (X_1, X_2, X_3, X_4) are independently drawn from uniform distributions: $X_{i1}, X_{i2}, X_{i3}, X_{i4} \overset{iid}{\sim}$ Unif$(-0.5, 0.5)$. The treatment indicator T is generated from $T_i \sim$ Bernoulli(p) with $p = \Lambda(0.5(X_{i1}, X_{i2}, X_{i3}, X_{i4}))$, where Λ denotes the logistic function. The potential outcomes $(Y^{(1)}, Y^{(0)})$ are generated according to: $Y_i = \mu^{(1)} D_i + \mu^{(0)}(1 - D_i) + (1 - \lambda) \cdot \epsilon_{1i} + \lambda \cdot \epsilon_{2i}$ where: $\mu^{(1)}(X_{i1}, X_{i2}, X_{i3}, X_{i4}) = 10 + 5X_{i1} + sin(10X_{i1}) - X_{i2} + 3tanh(0.5X_{i3})$, $\mu^{(0)}(X_{i1}, X_{i2}, X_{i3}, X_{i4}) = 0$ and the outlier indicator is generated as $\lambda \sim Bern(p)$ with $p \in \{0, 0.1, 0.2, 0.3, 0.4, 0.45\}$. We generate the non-contaminated error distribution term by $\epsilon_{1i} \overset{idd}{\sim} N(0, 0.25^2)$, and the contaminated error distribution by ϵ_{2i} with the following 4 different settings:

(a) Symmetric light-tail: $N(0, 10)$;

(b) Symmetric heavy-tail: $Cauchy(0, 0.5)$;

(c) Asymmetric mean 0: $logN(0, 1) - exp(1/2)$;

(d) Asymmetric mean $\neq 0$: Dirac Delta function with $\epsilon_{2i} = 20$.

The first two settings are used to assess the finite sample performance of our estimator under the assumption of symmetric error distributions. And the remaining two asymmetric settings are employed to evaluate the robustness of our estimators in scenarios where the assumption of symmetry is not met.

2.5.2 Verify The Asymptotic normality

For sample sizes $n \in \{500, 5000\}$ we estimate the CATE over $x_1 \in \{-0.4, -0.2, 0, 0.2, 0.4\}$ with 3 types of loss functions:

1. The L_2 loss;
2. The Huber loss function;
3. The Trukey's biweighted loss function.

In which use the L_2 loss function as an example of an unbounded loss function, the Hubers loss function as a bounded loss function, and the Tukeys biweighted loss function as examples of redescending loss functions.

For the choice of bandwidths, we use fixed bandwidths $h = 0.04$ for $n = 500$ and $h = 0.025$ for $n = 5000$. These values are obtained by averaging the bandwidths selected through cross-validation across repetitions.

With 1000 repetitions, we output the simulation bias, simulation variance, simulation MSE, mean of the estimated variance, and the probability of the 90% and 95% confidence interval constructed with estimated variances cover the truth.

From the simulation results, we observe the following:

Under symmetric cases, our proposed method exhibits similar bias, variance, and MSE compared to other methods when there are no outliers. In the presence of outliers, estimators based on redescending loss functions Tukey's biweighted loss outperform those based on Huber loss function and the non-robust estimator with L_2 loss. All estimators achieve coverage rates close to 95% and 90%, verifying both the asymptotic normality and the accuracy of our variance estimator.

In the asymmetric cases, the proposed method performs reasonably well with centered log-normal distributions, which are asymmetric with mean zero. In cases of asymmetric and non-centered means, while performance generally decreases, estimators with redescending loss functions Tukey's biweighted loss still maintain good performance even with $10\% - 20\%$ outliers.

Comparing $n = 500$ and $n = 5000$, we find that a larger sample size results in reduced bias, variance, and MSE, and improves the probability of achieving coverage rates closer to 95% and 90%. This indicates that the estimator approaches its theoretical normal distribution as the sample size increases.

2.5.3 Compare with Existing Methods

The estimators in Abrevaya, Hsu, and Lieli (2015) utilize higher-order kernels. To ensure a fair comparison with these estimators, we adopt a similar approach for our proposed

Table 2.1. Simulation results for robust CATE estimators using L_2, Huber, and Tukey's loss functions with $n = 500$ and outlier distribution $\mathcal{N}(0, 10)$.

	Bias			Var			MSE			EstVar			Rate95	
x0	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	Huber	Tukey
Contaminated Ratio=0														
-0.4	0.046	0.046	0.046	0.049	0.049	0.049	0.051	0.051	0.051	0.052	0.052	0.053	0.967	0.968
-0.2	0.048	0.048	0.047	0.040	0.040	0.040	0.042	0.042	0.043	0.051	0.051	0.052	0.978	0.977
0.0	-0.075	-0.075	-0.074	0.031	0.031	0.031	0.036	0.036	0.036	0.035	0.035	0.035	0.968	0.968
0.2	0.041	0.041	0.041	0.033	0.033	0.033	0.035	0.035	0.035	0.034	0.034	0.035	0.947	0.947
0.4	0.019	0.019	0.018	0.045	0.045	0.046	0.045	0.045	0.046	0.051	0.051	0.051	0.963	0.964
Contaminated Ratio=10%														
-0.4	0.047	0.050	0.055	0.648	0.262	0.106	0.650	0.265	0.109	0.678	0.285	0.116	0.982	0.972
-0.2	0.083	0.072	0.060	0.491	0.210	0.088	0.498	0.216	0.092	0.538	0.232	0.103	0.979	0.971
0.0	-0.105	-0.096	-0.086	0.425	0.176	0.069	0.436	0.185	0.076	0.439	0.183	0.076	0.970	0.966
0.2	0.045	0.036	0.028	0.406	0.175	0.069	0.408	0.176	0.069	0.454	0.191	0.078	0.978	0.964
0.4	0.005	0.008	0.009	0.509	0.226	0.096	0.509	0.226	0.096	0.552	0.249	0.107	0.983	0.973
Contaminated Ratio=20%														
-0.4	0.056	0.049	0.051	1.212	0.623	0.247	1.215	0.625	0.250	1.298	0.678	0.262	0.983	0.981
-0.2	-0.004	0.008	0.031	0.909	0.449	0.186	0.909	0.449	0.187	0.971	0.491	0.202	0.969	0.962
0.0	-0.076	-0.087	-0.083	0.820	0.397	0.155	0.826	0.404	0.162	0.824	0.411	0.161	0.973	0.966
0.2	0.045	0.033	0.025	0.790	0.389	0.142	0.792	0.390	0.143	0.859	0.432	0.166	0.974	0.971
0.4	0.071	0.043	0.019	1.072	0.534	0.209	1.077	0.535	0.210	1.039	0.528	0.217	0.967	0.967
Contaminated Ratio=30%														
-0.4	-0.025	-0.018	0.008	1.789	1.114	0.543	1.790	1.115	0.543	1.857	1.156	0.535	0.970	0.966
-0.2	0.033	0.041	0.052	1.376	0.828	0.383	1.377	0.829	0.386	1.446	0.880	0.412	0.975	0.974
0.0	-0.117	-0.105	-0.085	1.186	0.686	0.295	1.200	0.697	0.303	1.222	0.730	0.330	0.970	0.983
0.2	0.002	0.024	0.046	1.204	0.704	0.325	1.204	0.705	0.327	1.238	0.746	0.336	0.969	0.970
0.4	-0.024	-0.007	0.015	1.567	0.930	0.396	1.568	0.930	0.397	1.515	0.929	0.424	0.971	0.964
Contaminated Ratio = 40%														
-0.4	0.029	0.064	0.101	2.656	1.966	1.142	2.657	1.970	1.152	2.483	1.902	1.328	0.963	0.963
-0.2	0.028	0.034	0.048	1.820	1.303	0.742	1.820	1.304	0.745	1.898	1.398	0.803	0.969	0.969
0.0	-0.166	-0.140	-0.099	1.790	1.190	0.587	1.817	1.210	0.597	1.674	1.177	0.643	0.961	0.966
0.2	0.029	0.022	0.022	1.519	1.046	0.584	1.520	1.046	0.584	1.647	1.174	0.663	0.962	0.976
0.4	0.035	0.047	0.048	1.738	1.250	0.779	1.739	1.252	0.782	2.019	1.499	0.894	0.980	0.976
Contaminated Ratio = 45%														
-0.4	-0.029	-0.005	0.024	2.721	2.158	1.409	2.721	2.158	1.410	2.737	2.260	1.569	0.968	0.970
-0.2	0.023	0.038	0.045	2.033	1.512	0.925	2.034	1.514	0.927	2.144	1.690	1.087	0.962	0.974
0.0	-0.027	-0.029	-0.043	1.720	1.271	0.793	1.721	1.272	0.795	1.807	1.385	0.866	0.971	0.966
0.2	-0.009	0.009	0.025	1.765	1.332	0.780	1.765	1.332	0.781	1.898	1.470	0.900	0.970	0.966
0.4	0.018	0.013	-0.005	2.384	1.826	1.124	2.385	1.826	1.124	2.283	1.800	1.175	0.962	0.961

Table 2.2. Simulation results for robust CATE estimators with $n = 500$ and outlier distribution Cauchy$(0, 0.5)$.

	Bias			Var			MSE			EstVar			Rate95	
x0	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	Huber	Tukey
Contaminated Ratio=0														
-0.4	0.046	0.046	0.046	0.049	0.049	0.049	0.051	0.051	0.051	0.052	0.052	0.053	0.967	0.968
-0.2	0.048	0.048	0.047	0.040	0.040	0.040	0.042	0.042	0.043	0.051	0.051	0.052	0.978	0.977
0.0	-0.075	-0.075	-0.074	0.031	0.031	0.031	0.036	0.036	0.036	0.035	0.035	0.035	0.968	0.968
0.2	0.041	0.041	0.041	0.033	0.033	0.033	0.035	0.035	0.035	0.034	0.034	0.035	0.947	0.947
0.4	0.019	0.019	0.018	0.045	0.045	0.046	0.045	0.045	0.046	0.051	0.051	0.051	0.963	0.964
Contaminated Ratio=10%														
-0.4	0.038	0.049	4.30	1.43	0.074	0.059	4.30	0.076	0.061	1.73	0.074	0.061	0.967	0.964
-0.2	0.032	0.060	0.058	2.17	0.062	0.053	2.17	0.065	0.057	1.80	0.069	0.059	0.967	0.966
0.0	-0.051	-0.088	-0.086	7.00	0.046	0.037	7.00	0.053	0.045	7.10	0.051	0.041	0.972	0.969
0.2	4.22	0.032	0.032	1.83e2	0.047	0.038	1.83e2	0.048	0.039	1.28e2	0.050	0.041	0.966	0.964
0.4	-0.055	0.016	0.012	2.99	0.062	0.053	2.99	0.062	0.053	5.01	0.072	0.061	0.965	0.965
Contaminated Ratio=20%														
-0.4	1.58	0.056	0.048	1.14e3	0.097	0.069	1.14e3	0.100	0.071	9.90e2	0.096	0.069	0.976	0.970
-0.2	0.062	0.047	0.046	1.77	0.080	0.058	1.78	0.082	0.060	3.54	0.092	0.069	0.974	0.967
0.0	1.73	-0.072	-0.069	2.63e3	0.067	0.045	2.63e3	0.072	0.049	2.21e3	0.071	0.049	0.968	0.962
0.2	0.078	0.037	0.036	1.77	0.061	0.046	1.77	0.062	0.047	1.26	0.066	0.048	0.963	0.957
0.4	-0.037	0.043	0.043	6.28	0.093	0.066	6.28	0.095	0.068	6.35	0.096	0.070	0.976	0.973
Contaminated Ratio=30%														
-0.4	-0.073	0.054	0.054	8.67	0.121	0.077	8.67	0.124	0.080	4.60	0.122	0.082	0.973	0.973
-0.2	-2.14	0.063	0.062	2.41e3	0.091	0.063	2.42e3	0.095	0.067	3.08e3	0.105	0.075	0.979	0.975
0.0	0.411	-0.092	-0.090	1.14e2	0.077	0.048	1.14e2	0.085	0.056	3.24e2	0.088	0.057	0.982	0.971
0.2	0.161	0.039	0.038	1.57	0.080	0.055	1.57	0.081	0.057	7.04	0.083	0.056	0.964	0.955
0.4	0.105	0.033	0.029	3.94	0.104	0.072	3.94	0.105	0.073	4.98	0.114	0.078	0.976	0.967
Contaminated Ratio=40%														
-0.4	-0.143	0.057	0.058	1.385	0.163	0.092	1.387	0.166	0.095	1.337	0.159	0.096	0.974	0.962
-0.2	0.048	0.047	0.051	1.457e2	0.134	0.079	1.457e2	0.137	0.082	3.187e2	0.135	0.087	0.968	0.969
0.0	-0.219	-0.070	-0.072	2.748	0.108	0.067	2.753	0.113	0.072	5.205	0.103	0.064	0.963	0.956
0.2	-0.739	0.042	0.036	5.783e2	0.101	0.063	5.789e2	0.103	0.064	8.553e3	0.105	0.064	0.959	0.954
0.4	-0.406	0.026	0.026	8.671e2	0.123	0.071	8.673e2	0.124	0.072	1.103e4	0.135	0.088	0.972	0.969
Contaminated Ratio=45%														
-0.4	-0.581	0.044	0.048	2.021e2	0.165	0.092	2.024e2	0.167	0.094	2.571e2	0.177	0.103	0.975	0.963
-0.2	-0.278	0.049	0.047	7.050e2	0.124	0.080	7.051e2	0.126	0.082	8.709e2	0.141	0.092	0.977	0.970
0.0	-0.059	-0.074	-0.077	4.363	0.107	0.061	4.363	0.112	0.067	5.436	0.114	0.069	0.972	0.966
0.2	0.196	0.037	0.035	4.891	0.104	0.062	4.895	0.105	0.064	5.935	0.114	0.070	0.976	0.969
0.4	-0.226	0.011	0.005	6.353	0.145	0.087	6.358	0.145	0.087	5.164	0.155	0.096	0.970	0.966

Table 2.3. Simulation results for robust CATE estimators with $n = 500$ and outlier distribution $\log \mathcal{N}(0,1) - exp(1/2)$.

	Bias			Var			MSE			EstVar			Rate95	
x0	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	Huber	Tukey
Contaminated Ratio=0														
-0.4	0.046	0.046	0.046	0.049	0.049	0.049	0.051	0.051	0.051	0.052	0.052	0.053	0.967	0.968
-0.2	0.048	0.048	0.047	0.040	0.040	0.040	0.042	0.042	0.043	0.051	0.051	0.052	0.978	0.977
0.0	-0.075	-0.075	-0.074	0.031	0.031	0.031	0.036	0.036	0.036	0.035	0.035	0.035	0.968	0.968
0.2	0.041	0.041	0.041	0.033	0.033	0.033	0.035	0.035	0.035	0.034	0.034	0.035	0.947	0.947
0.4	0.019	0.019	0.018	0.045	0.045	0.046	0.045	0.045	0.046	0.051	0.051	0.051	0.963	0.964
Contaminated Ratio=10%														
-0.4	0.038	0.049	4.30	1.43	0.074	0.059	4.30	0.076	0.061	1.73	0.074	0.061	0.967	0.964
-0.2	0.032	0.060	0.058	2.17	0.062	0.053	2.17	0.065	0.057	1.80	0.069	0.059	0.967	0.966
0.0	-0.051	-0.088	-0.086	7.00	0.046	0.037	7.00	0.053	0.045	7.10	0.051	0.041	0.972	0.969
0.2	4.22	0.032	0.032	1.83e2	0.047	0.038	1.83e2	0.048	0.039	1.28e2	0.050	0.041	0.966	0.964
0.4	-0.055	0.016	0.012	2.99	0.062	0.053	2.99	0.062	0.053	5.01	0.072	0.061	0.965	0.965
Contaminated Ratio=20%														
-0.4	1.58	0.056	0.048	1.14e3	0.097	0.069	1.14e3	0.100	0.071	9.90e2	0.096	0.069	0.976	0.970
-0.2	0.062	0.047	0.046	1.77	0.080	0.058	1.78	0.082	0.060	3.54	0.092	0.069	0.974	0.967
0.0	1.73	-0.072	-0.069	2.63e3	0.067	0.045	2.63e3	0.072	0.049	2.21e3	0.071	0.049	0.968	0.962
0.2	0.078	0.037	0.036	1.77	0.061	0.046	1.77	0.062	0.047	1.26	0.066	0.048	0.963	0.957
0.4	-0.037	0.043	0.043	6.28	0.093	0.066	6.28	0.095	0.068	6.35	0.096	0.070	0.976	0.973
Contaminated Ratio=30%														
-0.4	-0.073	0.054	0.054	8.67	0.121	0.077	8.67	0.124	0.080	4.60	0.122	0.082	0.973	0.973
-0.2	-2.14	0.063	0.062	2.41e3	0.091	0.063	2.42e3	0.095	0.067	3.08e3	0.105	0.075	0.979	0.975
0.0	0.411	-0.092	-0.090	1.14e2	0.077	0.048	1.14e2	0.085	0.056	3.24e2	0.088	0.057	0.982	0.971
0.2	0.161	0.039	0.038	1.57	0.080	0.055	1.57	0.081	0.057	7.04	0.083	0.056	0.964	0.955
0.4	0.105	0.033	0.029	3.94	0.104	0.072	3.94	0.105	0.073	4.98	0.114	0.078	0.976	0.967
Contaminated Ratio=40%														
-0.4	0.107	-0.422	-0.669	2.914	0.580	0.360	2.925	0.758	0.807	2.744	0.579	0.430	0.976	0.960
-0.2	0.065	-0.476	-0.720	1.655	0.376	0.259	1.659	0.602	0.777	1.872	0.426	0.318	0.978	0.977
0.0	-0.084	-0.586	-0.822	1.512	0.343	0.234	1.519	0.686	0.909	1.596	0.369	0.273	0.970	0.961
0.2	0.003	-0.478	-0.705	1.585	0.300	0.213	1.585	0.529	0.710	1.621	0.366	0.270	0.975	0.963
0.4	0.015	-0.485	-0.731	1.763	0.466	0.299	1.763	0.702	0.834	1.760	0.463	0.342	0.965	0.961
Contaminated Ratio=45%														
-0.4	0.044	-0.523	-0.805	2.442	0.568	0.404	2.444	0.842	1.052	2.724	0.633	0.480	0.975	0.965
-0.2	0.123	-0.493	-0.799	2.348	0.427	0.287	2.363	0.670	0.925	2.356	0.497	0.365	0.967	0.966
0.0	-0.092	-0.646	-0.934	1.576	0.357	0.241	1.585	0.775	1.113	1.624	0.427	0.320	0.973	0.967
0.2	-0.019	-0.529	-0.796	1.428	0.372	0.258	1.429	0.652	0.891	1.558	0.414	0.314	0.973	0.963
0.4	0.049	-0.553	-0.836	3.850	0.572	0.380	3.852	0.878	1.078	3.490	0.523	0.399	0.951	0.947

Table 2.4. Simulation results for robust CATE estimators with $n = 500$ and outlier distribution Dirac Delta at $\epsilon_{2i} = 20$.

	Bias			Var			MSE			EstVar			Rate95	
x0	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	Huber	Tukey
Contaminated Ratio=0														
-0.4	0.046	0.046	0.046	0.049	0.049	0.049	0.051	0.051	0.051	0.052	0.052	0.053	0.967	0.968
-0.2	0.048	0.048	0.047	0.040	0.040	0.040	0.042	0.042	0.043	0.051	0.051	0.052	0.978	0.977
0.0	-0.075	-0.075	-0.074	0.031	0.031	0.031	0.036	0.036	0.036	0.035	0.035	0.035	0.968	0.968
0.2	0.041	0.041	0.041	0.033	0.033	0.033	0.035	0.035	0.035	0.034	0.034	0.035	0.947	0.947
0.4	0.019	0.019	0.018	0.045	0.045	0.046	0.045	0.045	0.046	0.051	0.051	0.051	0.963	0.964
Contaminated Ratio=10%														
-0.4	2.053	0.943	0.051	2.496	0.696	0.054	6.710	1.586	0.056	2.486	0.550	0.057	0.799	0.965
-0.2	2.135	0.973	0.062	1.749	0.510	0.049	6.305	1.458	0.053	1.979	0.443	0.058	0.877	0.972
0.0	1.948	0.797	-0.082	1.400	0.388	0.035	5.194	1.023	0.041	1.609	0.346	0.039	0.869	0.974
0.2	2.064	0.918	0.031	1.469	0.430	0.034	5.728	1.272	0.035	1.686	0.362	0.038	0.874	0.961
0.4	2.053	0.908	0.014	1.819	0.525	0.047	6.033	1.349	0.047	2.045	0.452	0.058	0.865	0.973
Contaminated Ratio=20%														
-0.4	4.117	2.576	0.050	4.075	3.039	0.061	21.029	9.675	0.063	4.949	2.017	0.064	0.833	0.971
-0.2	4.036	2.438	0.056	2.738	1.796	0.055	19.026	7.740	0.058	3.712	1.422	0.067	0.853	0.979
0.0	3.937	2.349	-0.066	2.704	1.888	0.041	18.203	7.406	0.045	3.183	1.252	0.045	0.835	0.970
0.2	3.999	2.422	0.037	2.681	1.899	0.042	18.676	7.764	0.043	3.289	1.270	0.044	0.829	0.958
0.4	4.054	2.471	0.038	2.935	2.056	0.060	19.369	8.161	0.061	3.969	1.556	0.068	0.874	0.966
Contaminated Ratio=30%														
-0.4	6.035	5.245	1.442	5.169	6.813	8.242	41.592	34.322	10.321	7.251	6.651	129.810	0.829	0.380
-0.2	6.073	5.213	1.265	3.933	5.354	5.096	40.815	32.529	6.697	5.603	4.897	5.759	0.848	0.405
0.0	5.826	4.962	1.123	3.544	5.081	5.033	37.491	29.700	6.293	4.709	4.184	2.891	0.834	0.406
0.2	6.018	5.137	1.203	3.169	4.550	4.452	39.391	30.941	5.899	4.921	4.229	1.901	0.871	0.396
0.4	5.995	5.142	1.245	3.979	5.626	5.338	39.916	32.068	6.888	5.842	5.161	78.255	0.849	0.411
Contaminated Ratio=40%														
-0.4	8.05	8.02	6.89	6.02	6.21	1.37	7.08	7.05	6.12	9.94	1.97	1.03e2	0.959	0.766
-0.2	8.11	8.08	6.83	4.31	4.50	1.11	7.01	6.98	5.77	7.38	1.26	1.40e2	0.965	0.793
0.0	7.90	7.88	6.55	3.77	3.89	9.70	6.62	6.60	5.27	6.34	1.05	2.22e6	0.965	0.765
0.2	8.09	8.07	6.76	3.81	3.98	9.74	6.93	6.90	5.55	6.57	1.10	6.62e2	0.970	0.801
0.4	8.10	8.07	6.88	4.77	4.98	1.15	7.04	7.01	5.88	7.91	1.38	1.44e3	0.958	0.803
Contaminated Ratio=45%														
-0.4	9.06	9.06	8.77	6.19	6.20	9.47	8.83	8.83	8.63	1.09	1.37	1.37e2	0.963	0.910
-0.2	9.12	9.12	8.81	4.64	4.64	7.45	8.78	8.78	8.50	8.34	1.00	5.53	0.976	0.928
0.0	8.87	8.87	8.49	3.76	3.77	6.32	8.24	8.24	7.84	7.12	8.21	1.21	0.986	0.923
0.2	9.00	9.00	8.64	3.94	3.95	6.48	8.49	8.49	8.11	7.41	8.86	1.75e3	0.981	0.929
0.4	9.05	9.05	8.72	4.72	4.73	7.54	8.66	8.66	8.36	8.84	1.06	1.27e2	0.979	0.931

Table 2.5. Simulation results for robust CATE estimators using L_2, Huber, and Tukey's loss functions with $n = 5000$ and outlier distribution $\mathcal{N}(0, 10)$.

	Bias			Var			MSE			EstVar			Rate95	
x0	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	Huber	Tukey
Contaminated Ratio=0														
-0.4	0.024	0.024	0.024	0.007	0.007	0.007	0.008	0.008	0.008	0.008	0.008	0.008	0.954	0.952
-0.2	0.016	0.016	0.016	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.962	0.962
0.0	-0.033	-0.033	-0.033	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.005	0.005	0.952	0.952
0.2	0.015	0.015	0.015	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.005	0.005	0.943	0.945
0.4	0.008	0.008	0.008	0.006	0.006	0.006	0.006	0.006	0.007	0.007	0.007	0.007	0.956	0.956
Contaminated Ratio=10%														
-0.4	0.013	0.016	0.020	0.102	0.040	0.016	0.102	0.041	0.017	0.101	0.040	0.016	0.960	0.950
-0.2	0.020	0.020	0.020	0.073	0.029	0.012	0.073	0.030	0.013	0.075	0.031	0.013	0.964	0.948
0.0	-0.020	-0.025	-0.028	0.065	0.027	0.011	0.066	0.028	0.011	0.064	0.026	0.011	0.955	0.950
0.2	0.009	0.011	0.014	0.068	0.027	0.010	0.068	0.027	0.011	0.067	0.027	0.011	0.962	0.961
0.4	0.003	0.005	0.006	0.085	0.034	0.012	0.085	0.034	0.012	0.083	0.034	0.014	0.956	0.961
Contaminated Ratio=20%														
-0.4	0.019	0.023	0.026	0.190	0.092	0.034	0.191	0.093	0.035	0.190	0.090	0.034	0.948	0.957
-0.2	0.030	0.026	0.020	0.146	0.067	0.024	0.147	0.068	0.025	0.144	0.067	0.026	0.957	0.957
0.0	-0.022	-0.020	-0.023	0.121	0.058	0.022	0.121	0.059	0.023	0.127	0.060	0.022	0.957	0.950
0.2	0.013	0.016	0.017	0.128	0.061	0.022	0.129	0.061	0.022	0.127	0.060	0.022	0.950	0.949
0.4	0.013	0.007	0.006	0.154	0.072	0.028	0.154	0.072	0.028	0.153	0.073	0.028	0.952	0.955
Contaminated Ratio=30%														
-0.4	-0.005	0.005	0.013	0.273	0.156	0.065	0.273	0.156	0.065	0.288	0.162	0.068	0.962	0.957
-0.2	0.031	0.025	0.017	0.219	0.121	0.049	0.220	0.122	0.049	0.214	0.122	0.052	0.949	0.955
0.0	-0.037	-0.034	-0.033	0.172	0.099	0.042	0.173	0.101	0.043	0.184	0.105	0.044	0.958	0.958
0.2	-0.008	-0.002	0.007	0.191	0.116	0.050	0.191	0.116	0.050	0.191	0.109	0.046	0.942	0.947
0.4	0.021	0.012	0.005	0.220	0.125	0.053	0.221	0.125	0.053	0.231	0.132	0.056	0.954	0.958
Contaminated Ratio=40%														
-0.4	0.022	0.025	0.024	0.386	0.265	0.132	0.386	0.266	0.132	0.375	0.255	0.131	0.945	0.958
-0.2	0.001	0.010	0.018	0.283	0.191	0.098	0.283	0.191	0.098	0.281	0.191	0.098	0.957	0.954
0.0	-0.042	-0.039	-0.034	0.258	0.168	0.081	0.260	0.170	0.082	0.247	0.168	0.086	0.957	0.961
0.2	0.031	0.027	0.020	0.275	0.177	0.084	0.276	0.178	0.085	0.253	0.171	0.088	0.947	0.955
0.4	0.000	0.004	0.011	0.287	0.194	0.100	0.287	0.194	0.100	0.307	0.209	0.108	0.956	0.962
Contaminated Ratio=45%														
-0.4	-0.011	0.002	0.018	0.392	0.284	0.175	0.392	0.284	0.175	0.423	0.311	0.178	0.967	0.957
-0.2	0.035	0.031	0.024	0.313	0.225	0.129	0.314	0.226	0.130	0.317	0.232	0.134	0.957	0.957
0.0	-0.061	-0.055	-0.043	0.270	0.192	0.107	0.273	0.195	0.109	0.279	0.204	0.117	0.953	0.950
0.2	0.038	0.038	0.031	0.283	0.208	0.125	0.284	0.210	0.126	0.281	0.205	0.118	0.940	0.940
0.4	0.010	0.012	0.014	0.338	0.244	0.137	0.339	0.244	0.137	0.341	0.252	0.146	0.956	0.959

Table 2.6. Simulation results for robust CATE estimators with $n = 5000$ and outlier distribution Cauchy$(0, 0.5)$.

	Bias			Var			MSE			EstVar			Rate95	
x0	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	Huber	Tukey
Contaminated Ratio=0														
-0.4	0.024	0.024	0.024	0.007	0.007	0.007	0.008	0.008	0.008	0.008	0.008	0.008	0.954	0.952
-0.2	0.016	0.016	0.016	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.962	0.962
0.0	-0.033	-0.033	-0.033	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.005	0.005	0.952	0.952
0.2	0.015	0.015	0.015	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.005	0.005	0.943	0.945
0.4	0.008	0.008	0.008	0.006	0.006	0.006	0.006	0.006	0.007	0.007	0.007	0.007	0.956	0.956
Contaminated Ratio=10%														
-0.4	-0.234	0.024	0.022	64.9	0.011	0.009	65.0	0.012	0.010	65.9	0.011	0.009	0.957	0.952
-0.2	0.965	0.025	0.022	903	0.009	0.007	903	0.009	0.007	2650	0.009	0.008	0.967	0.967
0.0	0.053	-0.024	-0.025	8.62	0.008	0.006	8.62	0.008	0.006	9.27	0.007	0.006	0.953	0.952
0.2	-0.007	0.015	0.015	1.17	0.007	0.006	1.17	0.007	0.006	1.80	0.007	0.006	0.959	0.960
0.4	-0.096	0.007	0.005	2.89	0.009	0.007	2.90	0.009	0.007	5.36	0.010	0.008	0.973	0.971
Contaminated Ratio=20%														
-0.4	0.040	0.021	0.021	1.57	0.015	0.011	1.57	0.016	0.011	2.71	0.015	0.010	0.947	0.946
-0.2	0.117	0.013	0.015	5.64	0.011	0.008	5.66	0.011	0.008	313	0.012	0.009	0.973	0.963
0.0	-0.131	-0.031	-0.029	11.0	0.010	0.007	11.0	0.011	0.008	12.8	0.010	0.007	0.952	0.951
0.2	0.072	0.014	0.015	5.81	0.010	0.007	5.82	0.010	0.007	6.16	0.010	0.007	0.954	0.950
0.4	0.217	0.012	0.010	24.3	0.012	0.009	24.4	0.013	0.009	19.5	0.013	0.009	0.961	0.952
Contaminated Ratio=30%														
-0.4	0.048	0.016	0.014	40.3	0.019	0.012	40.3	0.019	0.012	63.7	0.018	0.012	0.958	0.949
-0.2	0.076	0.025	0.022	15.9	0.014	0.009	15.9	0.015	0.010	28.9	0.015	0.010	0.957	0.957
0.0	-0.056	-0.031	-0.031	2.60	0.012	0.008	2.60	0.013	0.009	5.28	0.012	0.008	0.947	0.955
0.2	0.095	0.012	0.014	7.23	0.012	0.008	7.24	0.012	0.008	11.8	0.013	0.008	0.960	0.955
0.4	-0.022	0.006	0.005	50.4	0.015	0.010	50.4	0.015	0.010	55.5	0.016	0.011	0.953	0.955
Contaminated Ratio=40%														
-0.4	0.211	0.015	0.015	3.33	0.023	0.013	3.34	0.024	0.013	3.82	0.023	0.014	0.953	0.959
-0.2	0.074	0.029	0.024	1.20	0.017	0.010	1.20	0.017	0.011	3.35	0.018	0.011	0.957	0.959
0.0	0.180	-0.034	-0.032	5.95	0.014	0.008	5.95	0.015	0.009	5.11	0.015	0.009	0.958	0.964
0.2	1.57	0.020	0.020	2.71e3	0.016	0.010	2.71e3	0.016	0.010	2.61e3	0.015	0.010	0.951	0.952
0.4	-0.042	0.005	0.008	7.86	0.019	0.012	7.86	0.019	0.012	1.12e2	0.019	0.012	0.957	0.952
Contaminated Ratio=45%														
-0.4	-0.001	0.019	0.018	1.74	0.022	0.013	1.74	0.022	0.013	2.94	0.025	0.015	0.974	0.964
-0.2	0.182	0.023	0.024	9.34	0.017	0.011	9.37	0.018	0.011	3.31	0.019	0.012	0.969	0.973
0.0	0.258	-0.036	-0.034	1.59e2	0.016	0.009	1.60e2	0.017	0.010	1.36e2	0.017	0.010	0.956	0.959
0.2	-0.619	0.013	0.015	2.42e2	0.016	0.010	2.42e2	0.017	0.010	2.38e2	0.017	0.010	0.960	0.952
0.4	0.039	0.008	0.009	3.49	0.020	0.012	3.49	0.020	0.012	3.44	0.021	0.013	0.960	0.960

Table 2.7. Simulation results for robust CATE estimators with $n = 5000$ and outlier distribution $\log \mathcal{N}(0,1) - exp(1/2)$.

	Bias			Var			MSE			EstVar			Rate95	
x0	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	Huber	Tukey
Contaminated Ratio=0														
-0.4	0.024	0.024	0.024	0.007	0.007	0.007	0.008	0.008	0.008	0.008	0.008	0.008	0.954	0.952
-0.2	0.016	0.016	0.016	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.962	0.962
0.0	-0.033	-0.033	-0.033	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.005	0.005	0.952	0.952
0.2	0.015	0.015	0.015	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.005	0.005	0.943	0.945
0.4	0.008	0.008	0.008	0.006	0.006	0.006	0.006	0.006	0.007	0.007	0.007	0.007	0.956	0.956
Contaminated Ratio=10%														
-0.4	0.015	-0.121	-0.149	0.087	0.022	0.016	0.088	0.037	0.038	0.091	0.024	0.016	0.958	0.950
-0.2	0.012	-0.121	-0.147	0.073	0.017	0.012	0.073	0.032	0.033	0.069	0.018	0.013	0.958	0.962
0.0	-0.027	-0.166	-0.193	0.068	0.015	0.011	0.069	0.043	0.048	0.068	0.016	0.011	0.950	0.950
0.2	0.014	-0.125	-0.152	0.066	0.016	0.010	0.066	0.031	0.034	0.068	0.016	0.011	0.950	0.961
0.4	0.010	-0.131	-0.160	0.073	0.017	0.012	0.073	0.034	0.037	0.076	0.020	0.014	0.966	0.961
Contaminated Ratio=20%														
-0.4	0.003	-0.265	-0.338	0.226	0.036	0.026	0.226	0.107	0.140	0.216	0.041	0.028	0.969	0.967
-0.2	0.010	-0.250	-0.329	0.121	0.030	0.020	0.121	0.093	0.128	0.126	0.031	0.021	0.953	0.943
0.0	-0.023	-0.300	-0.380	0.119	0.025	0.016	0.120	0.115	0.161	0.122	0.028	0.019	0.968	0.967
0.2	-0.014	-0.268	-0.340	0.116	0.027	0.018	0.117	0.099	0.133	0.114	0.027	0.019	0.959	0.958
0.4	-0.008	-0.269	-0.344	0.154	0.031	0.023	0.154	0.104	0.141	0.149	0.033	0.023	0.958	0.950
Contaminated Ratio=30%														
-0.4	-0.004	-0.388	-0.536	0.228	0.059	0.042	0.228	0.209	0.329	0.237	0.061	0.042	0.956	0.945
-0.2	0.041	-0.375	-0.526	0.219	0.044	0.029	0.221	0.184	0.306	0.230	0.047	0.032	0.965	0.962
0.0	-0.028	-0.427	-0.576	0.200	0.038	0.025	0.201	0.220	0.357	0.190	0.040	0.028	0.958	0.963
0.2	0.025	-0.386	-0.535	0.172	0.036	0.024	0.173	0.184	0.310	0.188	0.041	0.028	0.966	0.964
0.4	0.008	-0.386	-0.535	0.217	0.046	0.029	0.217	0.194	0.316	0.222	0.050	0.034	0.955	0.959
Contaminated Ratio=40%														
-0.4	0.022	-0.505	-0.736	0.604	0.082	0.054	0.605	0.337	0.595	0.511	0.082	0.059	0.957	0.960
-0.2	-0.001	-0.507	-0.729	0.262	0.058	0.038	0.262	0.314	0.571	0.258	0.062	0.045	0.954	0.964
0.0	-0.032	-0.556	-0.782	0.266	0.046	0.033	0.267	0.356	0.645	0.274	0.054	0.039	0.976	0.974
0.2	0.016	-0.513	-0.742	0.253	0.048	0.034	0.253	0.311	0.585	0.267	0.055	0.040	0.968	0.959
0.4	-0.004	-0.526	-0.755	0.336	0.064	0.043	0.336	0.340	0.613	0.323	0.067	0.049	0.953	0.962
Contaminated Ratio=45%														
-0.4	0.012	-0.567	-0.842	0.509	0.081	0.055	0.510	0.402	0.765	0.475	0.093	0.069	0.976	0.975
-0.2	0.016	-0.562	-0.839	0.299	0.061	0.042	0.299	0.376	0.745	0.298	0.071	0.052	0.968	0.971
0.0	-0.012	-0.599	-0.878	0.270	0.053	0.036	0.270	0.412	0.807	0.279	0.062	0.046	0.961	0.971
0.2	0.011	-0.561	-0.835	0.262	0.058	0.039	0.262	0.372	0.736	0.265	0.062	0.046	0.958	0.969
0.4	0.027	-0.567	-0.841	0.389	0.069	0.047	0.390	0.390	0.753	0.375	0.076	0.056	0.957	0.970

Table 2.8. Simulation results for robust CATE estimators with $n = 5000$ and outlier distribution Dirac Delta at $\epsilon_{2i} = 20$.

	Bias			Var			MSE			EstVar			Rate95	
x0	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	L_2	Huber	Tukey	Huber	Tukey
Contaminated Ratio=0														
-0.4	0.024	0.024	0.024	0.007	0.007	0.007	0.008	0.008	0.008	0.008	0.008	0.008	0.954	0.952
-0.2	0.016	0.016	0.016	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.962	0.962
0.0	-0.033	-0.033	-0.033	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.005	0.005	0.952	0.952
0.2	0.015	0.015	0.015	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.005	0.005	0.943	0.945
0.4	0.008	0.008	0.008	0.006	0.006	0.006	0.006	0.006	0.007	0.007	0.007	0.007	0.956	0.956
Contaminated Ratio=10%														
-0.4	2.019	0.856	0.019	0.308	0.076	0.009	4.386	0.809	0.009	0.378	0.073	0.008	0.940	0.950
-0.2	2.012	0.853	0.020	0.255	0.062	0.006	4.304	0.790	0.007	0.281	0.055	0.007	0.915	0.964
0.0	1.976	0.809	-0.028	0.211	0.053	0.006	4.115	0.707	0.006	0.247	0.048	0.006	0.927	0.956
0.2	2.024	0.854	0.014	0.203	0.051	0.005	4.299	0.781	0.005	0.255	0.050	0.006	0.945	0.960
0.4	2.015	0.847	0.005	0.281	0.070	0.006	4.341	0.788	0.006	0.306	0.060	0.008	0.911	0.974
Contaminated Ratio=20%														
-0.4	4.033	2.322	0.022	0.595	0.326	0.009	16.858	5.717	0.010	0.749	0.254	0.009	0.908	0.952
-0.2	3.998	2.290	0.017	0.457	0.255	0.007	16.437	5.499	0.007	0.556	0.188	0.008	0.888	0.965
0.0	3.985	2.265	-0.027	0.376	0.219	0.006	16.254	5.347	0.007	0.491	0.167	0.006	0.913	0.956
0.2	3.993	2.285	0.015	0.417	0.237	0.006	16.359	5.457	0.006	0.500	0.168	0.006	0.896	0.960
0.4	3.989	2.281	0.006	0.485	0.282	0.008	16.394	5.486	0.008	0.600	0.204	0.009	0.892	0.949
Contaminated Ratio=30%														
-0.4	6.002	5.009	0.474	0.832	1.252	0.139	36.857	26.339	0.363	1.112	0.789	0.030	0.869	0.688
-0.2	6.032	5.020	0.479	0.535	0.844	0.115	36.917	26.047	0.344	0.838	0.589	0.024	0.896	0.642
0.0	5.960	4.943	0.417	0.490	0.795	0.110	36.014	25.230	0.284	0.726	0.511	0.020	0.880	0.587
0.2	6.052	5.050	0.478	0.524	0.877	0.129	37.152	26.380	0.357	0.756	0.534	0.020	0.870	0.557
0.4	5.978	4.972	0.470	0.647	1.030	0.133	36.384	25.750	0.354	0.900	0.634	0.026	0.863	0.597
Contaminated Ratio=40%														
-0.4	8.011	8.011	6.648	0.904	0.904	2.368	65.073	65.072	46.568	1.481	2.980	2.005	0.996	0.875
-0.2	8.002	8.002	6.612	0.662	0.662	1.770	64.698	64.698	45.490	1.111	2.192	1.293	0.997	0.859
0.0	7.949	7.949	6.550	0.615	0.615	1.705	63.810	63.810	44.613	0.969	1.904	1.105	0.995	0.862
0.2	8.021	8.021	6.633	0.574	0.574	1.623	64.911	64.911	45.622	0.999	1.987	1.147	0.996	0.872
0.4	8.052	8.052	6.698	0.731	0.731	1.982	65.559	65.559	46.848	1.208	2.408	1.468	0.993	0.882
Contaminated Ratio=45%														
-0.4	9.031	9.031	8.776	0.930	0.930	1.467	82.495	82.495	78.484	1.672	1.673	2.337	0.991	0.970
-0.2	9.030	9.030	8.768	0.668	0.668	1.066	82.202	82.202	77.952	1.253	1.255	1.727	0.991	0.967
0.0	8.980	8.980	8.720	0.586	0.586	0.940	81.232	81.232	76.978	1.093	1.094	1.497	0.993	0.976
0.2	9.009	9.009	8.743	0.652	0.652	1.045	81.811	81.811	77.483	1.125	1.126	1.539	0.990	0.972
0.4	9.023	9.023	8.765	0.742	0.742	1.182	82.163	82.163	78.006	1.352	1.354	1.869	0.988	0.969

estimators by using sixth-order kernels, constructed as described in G. W. Imbens and Ridder (2009). The bandwidths are selected through cross-validation for each method in each repetition.

For the comparison, we estimate the CATE using each of the methods at 100 equidistant points between -0.5 and 0.5, with 1000 repetitions. The MISE, interval width, and coverage probability of confidence intervals are computed and summarized in Figure 2.1.

Our comparison reveals that, in non-contaminated cases, our proposed estimators perform comparably to existing methods. However, in the presence of outliers, our robust loss function-based methods consistently outperform the existing non-robust estimators.

2.6 Real Data Application

The application involves data from the 2007-2008 NHANES, a publicly available dataset provided by the United States Centers for Disease Control and Prevention (CDC). The goal is to estimate the effect of physical activity on albumin level changes with age, while controlling for other confounders.

Albumin levels are obtained from laboratory data, measured in milligrams of albumin per gram of creatinine (mg/g). Given the right-skewed nature of the data and the lack of observations close to 0 (unlike Robinson et al. (2010)), we apply a log transformation to the response variable Y. Despite the transformation, the scatter plot indicates that the error distribution remains slightly asymmetric, with some observations deviating significantly from the majority. This motivates the use of our robust method for this dataset. Our simulation results suggest that our proposed method performs well under slightly skewed and low contamination conditions.

The treatment indicator is derived from the physical activity questionnaire (prefix PAQ), specifically using the variable "vigorous work activity" as the treatment indicator T. We select observations with $T = 1$ and $T = 0$ for our estimation and exclude other values.

The covariates X include Age, Ratio of family income to poverty (from 0 to 5), Gender, Race, Education Level, BMI, diabetes, history of kidney disease, and blood pressure. These variables are commonly used as confounders or in experimental designs (Chang et al., 2013;

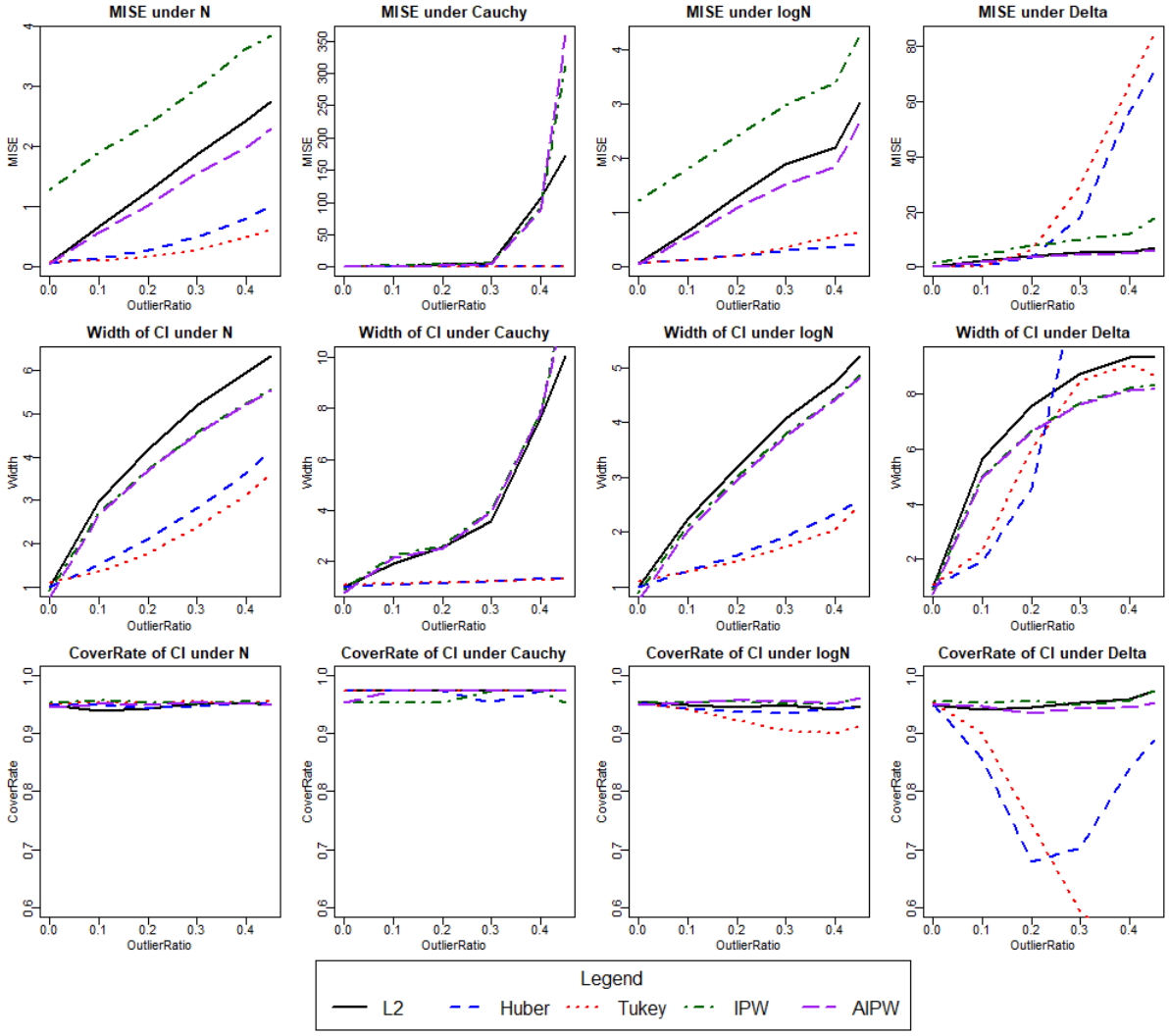

Figure 2.1. Comparison of MISE, width of confidence intervals, and coverage probability between the proposed method, L2, Huber, and Tukey's loss functions, as well as IPW and AIPW methods, across varying error distributions and outlier ratios.

Lang et al., 2018). We assume unconfoundedness given this set of covariates and use age as the conditioned covariate.

After merging the dataset and removing observations with missing values, the final sample size is $n = 5021$, with 982 observations in the treatment group and 4039 in the control group.

Elevated urinary albumin excretion is a predictor of future cardiovascular disease, hypertension, and chronic kidney disease (Robinson et al., 2010; Comper and Osicka, 2005; Solbu et al., 2008; W. G. Miller et al., 2009). Physical activity is often negatively associated with albuminuria, indicating potential protective effects on cardiovascular and renal health (Kuo et al., 2022). In diabetic populations, physical activity is linked to lower albumin excretion, and interventions have shown promise in reducing albuminuria (Hambrecht, Fiehn, et al., 1998; Hambrecht, Wolf, et al., 2000; Leon et al., 1987).

However, findings in non-diabetic individuals are mixed. Some studies have found a significant association between physical activity and reduced albumin excretion (Robinson et al., 2010), while others have reported no such association (Chang et al., 2013). These inconsistencies highlight the need to explore the heterogeneity of the effect across different subpopulations, a topic that has not been extensively studied in the literature.

These mixed findings underscore the importance of investigating how physical activity impacts albuminuria across various subpopulations, particularly considering continuous covariates that might influence this relationship.

Compared to the IPW and AIPW methods, the overall trend and value of our CATE estimators are similar to the doubly robust estimator. The estimated functions are in general all lower than 0, which is consistent with the existing results (Kuo et al., 2022; Hambrecht, Fiehn, et al., 1998; Hambrecht, Wolf, et al., 2000; Leon et al., 1987). However, the widths of the confidence interval of our estimator with robust loss functions (Huber's and Tukey's) are smaller and hence we're able to conclude that with age over 60, we're 95% confident that vigorous work activities decrease the $log(Albumin)$ value. Furthermore, with the redescending loss function the Tukey's biweighted loss, the estimated functions are close to 0 before age 45 without an obvious increasing trend. While in the Huber loss and other nonrobust methods, the estimated functions are negative with an increasing trend before 40 years old, which is not supported by existing studies.

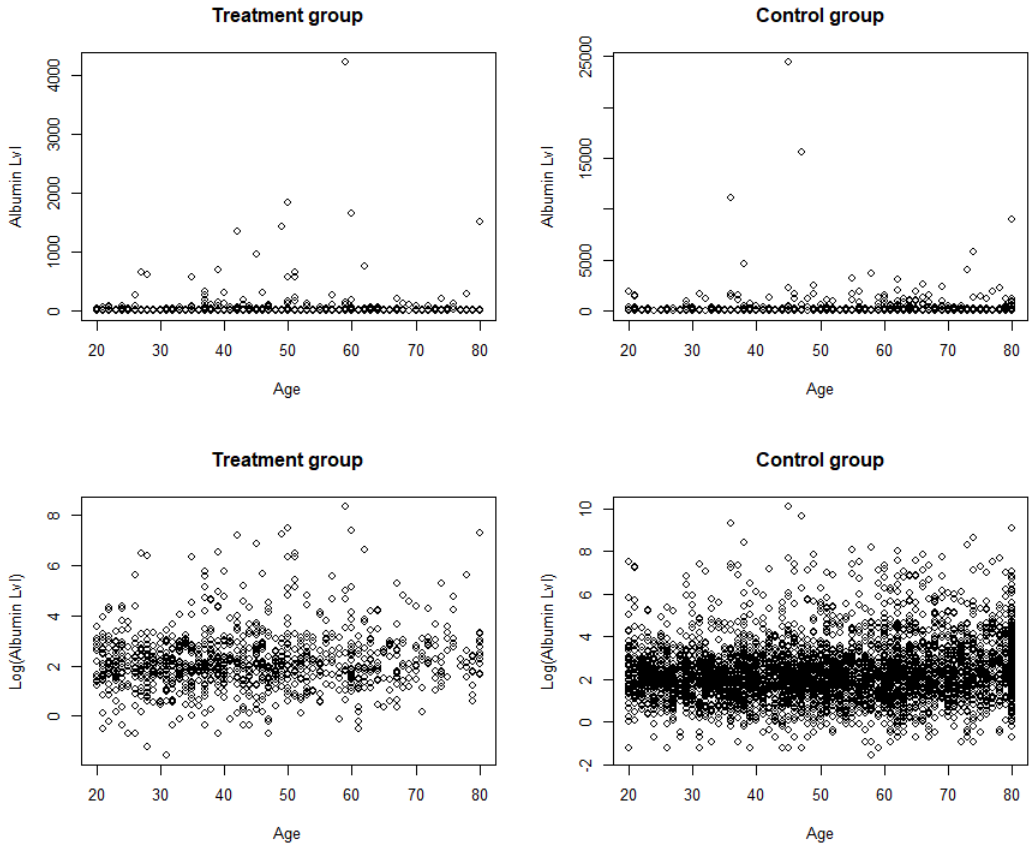

Figure 2.2. Scatter plot of Albumin levels conditioned on age, and scatter plot of log-transformed Albumin levels conditioned on age, separated by whether the individual engages in vigorous work activities.

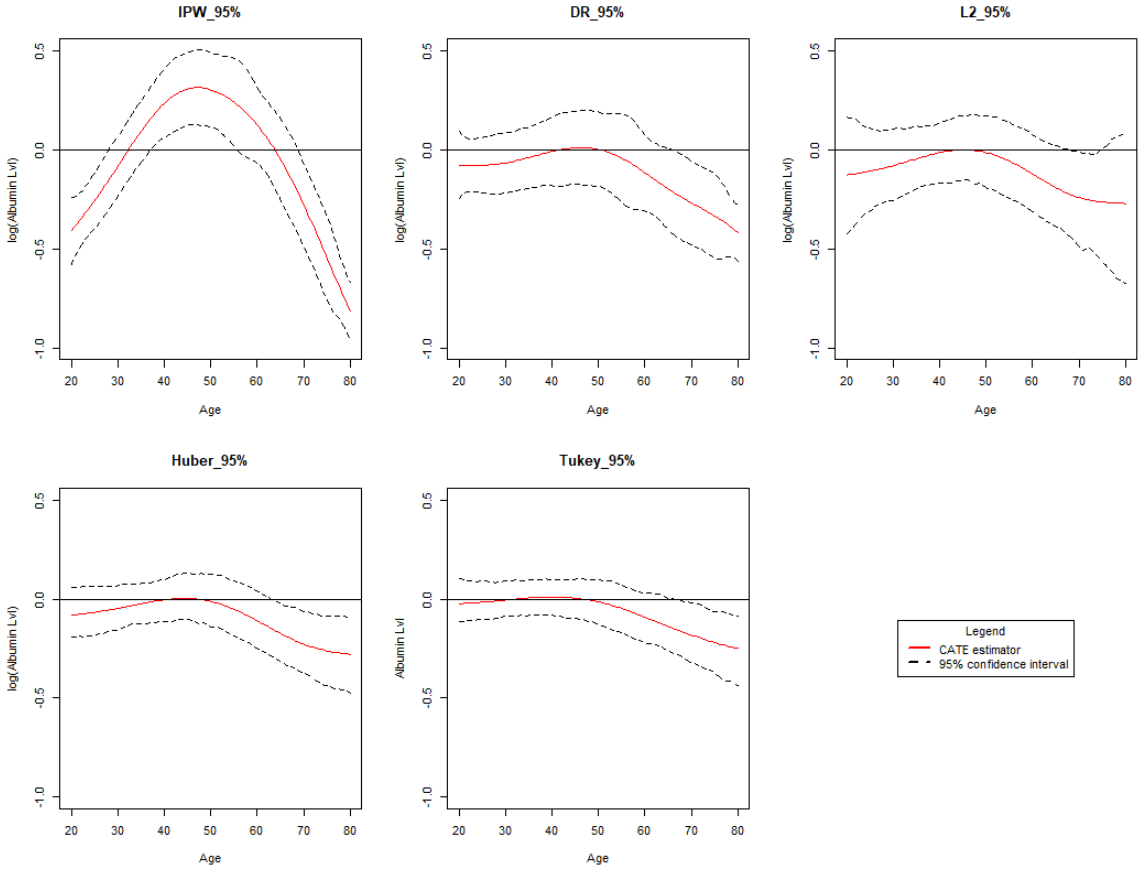

Figure 2.3. CATE of the effect of vigorous work activity on the log-transformed urinary albumin levels, conditioned on age, along with the 95% confidence intervals.

2.7 Discussion

In this paper, we propose an outlier-resistant Inverse Probability Weighting (IPW) method for estimating the CATE. We address the scenario where a high-dimensional vector of covariates is required to identify the CATE, but the covariates of primary interest are of much lower dimension.

Our proposed estimators are shown to be consistent and asymptotically normal, demonstrating strong finite sample performance, which ensures valid inference and prediction. When the sample distribution is not contaminated, our method performs similarly to IPW and Doubly Robust (DR) estimators, though with slightly lower efficiency. In the presence of contamination, our method outperforms both IPW and DR methods. We establish the asymptotic properties of our estimators and present a pointwise confidence interval, illustrating their utility through Monte Carlo experiments and an application to the effects of smoking on birth weights.

Several directions for future research are evident. First, exploring the impact of asymmetric density functions $f(y \mid x)$ could provide additional insights. Second, model misspecification issues may arise since our approach relies on a parametric method to estimate $\boldsymbol{\pi}(X)$. Third, developing a fully nonparametric method could potentially improve efficiency, as indicated by previous work (Abrevaya, Hsu, and Lieli, 2015; Hirano, G. W. Imbens, and Ridder, 2003). Finally, given the high-dimensional nature of covariates X, incorporating techniques such as variable selection or dimension reduction could enhance the method's performance.

3. OUTLIER RESISTANT INFERENCE FOR HETEROGENEOUS TREATMENT EFFECT IN THE ABSENCE OF SYMMETRY AND LIGHT TAIL ASSUMPTIONS

The CATE is essential for understanding treatment heterogeneity, but traditional estimation methods often struggle with outliers and heavy-tailed errors. To address these challenges, we propose a robust estimator for CATE that does not rely on assumptions of symmetry or light-tailed distributions. Unlike conventional M-estimators, which assume symmetric error distributions, our approach utilizes the adaptive Huber loss, which adjusts its robustification parameter according to sample size, providing both asymptotic unbiasedness and robustness even without the symmetric or light-tailed distribution assumptions. Procedures for selecting the optimal robustification parameter and constructing asymptotically valid confidence intervals are developed based on our theoretical results. The method is validated through simulations and applied to NHANES data to investigate the impact of heavy alcohol consumption on liver enzyme levels across different ages.

3.1 Introduction

Estimating the CATE offers valuable insights into treatment effects across diverse subpopulations. Under the counterfactual framework (Rubin, 1974; Neyman, 1923), Abrevaya, Hsu, and Lieli (2015) proposed an inverse propensity weighting (IPW) estimator for situations where the unconfoundedness assumption does not hold generally, conditional on low-dimensional covariates. The IPW method operates in a two-step framework: first, estimating the propensity score, and then using this estimate to construct a target function. The conditional mean of this target function is then estimated, conditioned on a subset of the covariates. Since this subset is typically assumed to be low-dimensional, kernel smoothing is employed in the second step.

However, kernel smoothing based on the L_2 loss is sensitive to outliers and exhibits large variance when dealing with heavy-tailed error distributions. Consequently, the performance of these methods is limited, necessitating the development of an outlier-resistant approach.

In real-world applications, data are often contaminated with outliers or contain heavy-tailed variables, making many conventional methods based on squared loss inadequate. To address this challenge, traditional robust regression methods using M-estimators (Huber, 1964) assume that the noise distribution is symmetric and down weight the influence of outliers through specific loss functions. One notable example is Huber loss (Huber, 1964), which combines characteristics of both the L_1 (absolute error) and the L_2 (squared error) loss functions, offering a compromise between their properties. When the error distribution is symmetric, Härdle (1984) extends the work of Huber (1964) by combining kernel regression and M-estimation to create a robust local estimator, deriving the asymptotic properties of the local M-estimator.

Following Abrevaya, Hsu, and Lieli (2015) and Härdle (1984), we derived a robust estimator of the CATE with symmetric assumptions in Chapter 2. In the presence of asymmetric errors, the traditional M-estiamtor that rely on the assumption of symmetric errors may no longer perform effectively in general. For example, the conditional median estimated using the L_1 loss can differ significantly from the conditional mean in models with asymmetric errors. Consequently, the proposed outlier-resistant estimator based on the traditional M-estimator for the CATE in Chapter 2, following Abrevaya, Hsu, and Lieli (2015) and Härdle (1984), is generally biased. To the best of our knowledge, no existing literature addresses the estimation of CATE under heavy-tailed and asymmetric potential outcomes. Therefore, this paper focuses on developing methods for estimating CATE in the presence of such heavy-tailed and asymmetric error conditions.

To address the challenge of estimating the conditional mean in the presence of asymmetric and heavy-tailed error distributions, we reviewed various methods but did not apply them for specific reasons. To develop a robust and unbiased estimator, researchers often resort to transformation techniques in regression analysis. Examples include logarithmic or Box-Cox transformations (Box and Cox, 1964). However, these methods introduce a significant bias known as transformation bias to the estimated conditional mean. Despite efforts

to mitigate this bias (Duan, 1983; D. M. Miller, 1984), the determination of an appropriate transformation parameter often lacks stability, leading to instability in the resulting conditional mean estimates (Bickel and Doksum, 1981). Except for the transformation methods, researchers have explored alternative approaches by analyzing residual characteristics to develop robust and unbiased conditional mean estimators. For instance, Moberg, Ramberg, and Randles (1978) proposed an adaptive M-estimator for location estimation problems based on a lambda distribution with three parameters. Similarly, Fu and Y.-G. Wang (2021) constructed an asymmetric Tukeys bi-weighted loss function with two tuning parameters and developed a data-driven method to identify the most appropriate tuning parameters by minimizing the variance estimator. Additionally, D. Wang, Romagnoli, and Safavi (2000) introduced a wavelet-based robust M-estimation method that constructs a robust loss function by estimating the error distribution nonparametrically. Although these methods aim to provide unbiased and robust estimators, they generally assume a consistent error distribution across all values of the conditioned variables. When there is variation in the error distribution, accurately determining the tuning parameters for the robust loss function can become impractical. Additionally, Takeuchi, Bengio, and Kanamori (2002) and Kanamori and Takeuchi (2006) proposed a method by estimating a sequence of conditional quantile estimators and then fitting a second regression to obtain the conditional mean. However, as noted by Burke et al. (2019), building a number of quantile regressions is not preferred in practice due to its complexity and computational demands.

To provide a method for handling heavy-tailed asymmetric errors with high-dimensional covariates, J. Fan, Q. Li, and Y. Wang (2017) proposed a penalized Huber loss with a parameter that increases as the sample size grows. The key observation is that the robustification parameter should diverge at a certain rate to achieve an optimal trade-off between bias and robustness (Sun, W.-X. Zhou, and J. Fan, 2020). They obtained a robust estimator that is concentrated around the true mean with exponentially high probability. Sun, W.-X. Zhou, and J. Fan (2020) extended this result, requiring only the existence of the $1 + \delta$ moment for any $\delta > 0$. Additionally, W.-X. Zhou et al. (2018) focused on the large-scale multiple testing problem, providing the Berry-Esseen inequality and Cramer-type moderate deviation results that show how the estimator converges to a normal distribution. J. Luo, Sun, and W.-X.

Zhou (2022) demonstrated how these results can be applied to distributed data, using the Berry-Esseen type inequality as theoretical evidence for constructing confidence intervals. Unlike the other methods we've reviewed, these approaches offer promising solutions to our problem, as they do not rely on assumptions of symmetry or homoscedasticity, but only require the existence of certain moments.

In this paper, we adopt the adaptive Huber regression for the estimation of the CATE to ensure both asymptotic unbiasedness and robustness in the presence of heavy-tailed and asymmetrically distributed error terms. To establish the theoretical properties of the proposed estimator, we derive a concentration inequality and a Berry-Esseen type bound. Our proposed estimator is consistent, does not require the assumption of symmetric errors, and only necessitates the existence of the second moment for consistency and the third moment for asymptotic normality. At the same time, it offers a level of robustness compared to non-robust methods based on the squared loss. Additionally, since the concentration inequality is too conservative for inference purposes, we propose a normal-based asymptotically correct confidence interval using the Berry-Esseen bound for practical application. Procedures for selecting the robustification parameter and constructing asymptotically valid confidence intervals are provided based on the theoretical results.

Furthermore, we apply our proposed method to examine the CATE of alcohol consumption on liver function biomarkers across different age groups using data from the NHANES for 1999-2006. The biomarkers of interestalanine aminotransferase (ALT), aspartate aminotransferase (AST), gamma-glutamyl transferase (GGT), and alkaline phosphatase (ALP)are widely recognized indicators of liver function and are commonly influenced by alcohol intake. Existing studies suggest that heavy drinking leads to elevated levels of these biomarkers, though the effect may vary by age (Torkadi, Apte, and Bhute, 2014; Lala, Zubair, and Minter, 2023; Conigrave et al., 2003; Alatalo et al., 2009; Hietala et al., 2005; Agarwal, Fulgoni, and Lieberman, 2015). Our proposed robust CATE estimator, using adaptive Huber loss, demonstrates strong resistance to outliers while avoiding the irreducible bias common in standard robust methods. By capturing the heterogeneity in the data, our method provides a more refined understanding of how alcohol affects liver function across the lifespan. Our results show that heavy drinking has a positive effect on all four biomarkers, aligning with

existing studies, and reveals a greater effect in middle-aged individuals compared to younger or older groups. Possible explanations and further insights are discussed in Section 7.

The rest of the paper is organized as follows: Section 2 introduces the CATE estimator with adaptive Huber loss, covering its identification and estimation. Section 3 develops the theoretical properties of the proposed estimators, while Section 4 derives confidence intervals based on these results. Section 5 presents the algorithm for the proposed method. Section 6 discusses the simulation results, followed by the data analysis in Section 7. Finally, Section 8 concludes the paper.

3.2 Methodology

3.2.1 Problem Setup

To describe our methodology, we first outline the commonly used potential outcome framework (Rubin, 1974; Neyman, 1923). Let $Y = TY^{(1)} + (1-T)Y^{(0)}$ where $Y^{(1)}$ and $Y^{(0)}$ represent the potential outcomes with and without treatment, respectively, corresponding to the treatment indicator $T = 1$ and $T = 0$. We observe n independent and identically distributed copies $\{(Y_{\mathrm{i}}, T_{\mathrm{i}}, X_{\mathrm{i}}) : \mathrm{i} = 1...n\}$ from the joint distribution (Y, T, X), where $X \in R^p$ denotes a p-dimensional vector of covariates. In this setup, the CATE is defined as

$$\tau(x_1) = E[Y^{(1)} - Y^{(0)} \mid X_1 = x_1],$$

where $X_1 \in R^d$ is a d-dimensional subvector of X. While our results apply to both $d = p$ and $d \ll p$, we focus particularly on the latter case.

Following Chapter 2 we estimate $\mu_1(x_1) = E[Y^{(1)} \mid X_1 = x_1]$ and $\mu_0(x_1) = E[Y^{(0)} \mid X_1 = x_1]$ separately, and then estimate the CATE by a subtraction $\hat{\tau}(x_1) = \hat{\mu}_1(x_1) - \hat{\mu}_0(x_1)$. For simplicity, we focus on the estimation of $\hat{\mu}_1(x_1)$ since $\hat{\mu}_0(x_1)$ can be estimated similarly.

Since potential outcomes $Y^{(1)}$ and $Y^{(0)}$ are not observed simultaneously, we rely on the unconfoundedness assumption (Rosenbaum and Rubin, 1983) and the positivity assumption adopted in Abrevaya, Hsu, and Lieli (2015) to identify the CATE:

Assumption (A1)

1. (Unconfoundedness) $(Y^{(1)}, Y^{(0)}) \perp T \mid X$.

2. (Positivity) Let $\boldsymbol{\pi}(x) = P(T = 1 \mid X = x)$ be the propensity score. There exists $C > 0$ s.t. $P(C \leq \boldsymbol{\pi}(X) \leq 1 - C) = 1$.

Under these assumptions, the conditional mean $\mu_1(x_1)$ from Chapter 2 is determined by solving the estimating equation:

$$0 = E\left[\frac{T}{\boldsymbol{\pi}(X)}(Y - \mu_1(x_1)) \mid X_1 = x_1\right].$$

We then modify this equation using a robust ψ function:

$$0 = E\left[\frac{T}{\boldsymbol{\pi}(X)}\psi(Y - \mu_1(x_1)) \mid X_1 = x_1\right].$$

In Chapter 2, we demonstrated that if the conditional distribution $f(y^{(1)} \mid x_1)$ is symmetric and ψ is anti-symmetric, the modified estimating equation remains unbiased, meaning that its solution is equivalent to the original unmodified equation. However, this result generally does not hold without the symmetry assumption.

3.2.2 Adaptive Huber loss

The robust estimator from Chapter 2 is defined as the solution to the empirical form of the estimating equation:

$$0 = \sum_{\mathrm{i}=1}^{n} \frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}\psi(Y_{\mathrm{i}} - \hat{\mu}_1(x_1))K_h(X_{\mathrm{i}1} - x_1),$$

where $\hat{\boldsymbol{\pi}}(x)$ is the estimated propensity score and $K_h(x) = \frac{1}{h}K(\frac{x}{h})$ with K as a kernel function with order s. Asymptotic results from Chapter 2 show that the estimator's bias takes the form:

$$B_n(x_1) = E[\psi(Y^{(1)} - \mu_1(x_1)) \mid X_1 = x_1]g(x_1) + O(h^s),$$

where the first term represents the bias due to the robust loss function, and the second term accounts for the bias introduced by kernel smoothing. As h approaches zero with increasing sample size, the second term diminishes. However, without the unbiasedness of the estimating equation, the first term does not generally equal to or approach to zero, indicating that the bias of the estimator cannot always diminish as the sample size increases.

Following the approach of J. Fan, Q. Li, and Y. Wang (2017), we propose using the Huber loss with an adaptive robustification parameter to achieve both robustness and asymptotic unbiasedness. Adopting the terminology of Sun, W.-X. Zhou, and J. Fan (2020), we refer to this approach as the adaptive Huber loss.

$$\rho_\alpha(u) = \begin{cases} u^2/2, & \mid u \mid \leq \alpha, \\ \alpha \mid x \mid -\alpha^2/2, & \mid u \mid > \alpha, \end{cases}$$

where $\alpha > 0$ is referred to as the robustification parameter, which balances unbiasedness and robustness (J. Fan, Q. Li, and Y. Wang, 2017). Similar to the original Huber loss, for any specific value of α, the loss function is quadratic for small values of x and linear for large values. However, unlike the classical Huber estimator (Huber, 1973), where α is fixed, we allow α to be determined by the sample size. Specifically, as $\alpha \to 0$, the loss function becomes the L_1 loss, which is robust but biased. Conversely, as $\alpha \to \infty$, the loss function becomes the L_2 loss, which is not robust but unbiased. With a small sample, a single gross outlier can significantly deviate the estimator, making robustness more crucial. In contrast, with a large sample, the presence of enough outliers can help capture the error distribution accurately, making unbiasedness more critical. The L_2 loss is highly sensitive to outliers due to its quadratic nature, resulting in large variance, particularly in small samples (Huber, 1973; Catoni, 2012). The standard Huber loss, with a fixed α, may introduce non-negligible estimation bias (Sun, W.-X. Zhou, and J. Fan, 2020), making it unsuitable for asymmetrically distributed errors, especially in large samples.

To address this drawback, the adaptive Huber loss, with a suitably chosen α adptive based on the sample size, achieves asymptotically unbiasedness comparing to the fixed Huber loss while still keeping a certain level of robustness comparing to the L_2 loss.

By the definition of ρ_α, we define ψ_α as the derivative of the adaptive loss function:

$$\psi_\alpha(u) = \begin{cases} u, & \mid u \mid \leq \alpha, \\ -1, & u < -\alpha, \\ 1, & u > \alpha. \end{cases}$$

Then, we replace the function ψ with the derivative of adaptive Huber loss function ψ_α and define our estimator $\hat{\mu}_1(x_1)$ as the solution of the estimating equation:

$$0 = \sum_{\mathrm{i}}^{n} \frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})} \psi_\alpha(Y_{\mathrm{i}} - \hat{\mu}_1(x_1)) K_h(X_{\mathrm{i}1} - x_1).$$

When the adaptive robustification parameter α is determined, we may similarly determine the bandwidths and employ the iteratively reweighted least square(IRLS) method for estimation. The details of selecting α will be discussed in a later section, as the theoretical results are necessary for its proper determination.

3.3 Theoretical Result

We start with the following regularity conditions.

In this paper, our primary focus is on the use of adaptive Huber loss. To simplify our analysis, we assume that the propensity score is estimated using a parametric method with a sufficiently fast convergence rate, as detailed in (A2).

(A2) (Estimated Propensity score) Follow the semiparametric method in Abrevaya, Hsu, and Lieli (2015), we assume the propensity score is correctly specified and estimated by parametric method with $sup_{x \in \mathcal{X}} \mid \hat{\boldsymbol{\pi}}(X) - \boldsymbol{\pi}(X) \mid = O_p(1/\sqrt{n})$. Which is typically hold for standard parametric estimation methods under reasonably mild regularity conditions.

(A3) (Distributions of X and Y) Assume the conditional probability density function $f(Y^{(1)} \mid X_1)$ has a bounded partial derivative with respect to $x_1 \in \mathcal{X}$, where $\mathcal{X}$ is a Cartesian product of compact intervals on the real line. The density function $g(x_1)$ of X_1 is assumed to be positive on its support and twice differentiable.

(A4) (Conditional moments and smoothness) The conditional second moments $\sup_{x_1 \in \mathcal{X}} E[Y^{(\mathrm{j})2} \mid X_1 = x_1] < \infty$ for $\mathrm{j} = 0, 1$. The functions $m_{\mathrm{j}}(x) = E[\psi(Y^{(\mathrm{j})} - \mu_{\mathrm{j}}) \mid X = x], \mathrm{j} = 0, 1$ and $g(x_1)$ the density function of X_1 are s times continuously differentiable on $x \in \mathcal{X}$.

(A5) (Kernel)$K(u)$ is a kernel of order $s \geq 2$ symmetric with respect to 0.

(A6) (Bandwith) The bandwidths h satisfy the conditions: $nh^{2s+d} \to 0$ and $nh^d \to \infty$ as $n \to \infty$.

Same as Chapter 2, we apply the result from Härdle (1984) based on the following definition of the delta function sequence(DFS) denoted by $\delta_n(\cdot)$:

(D1) $\int \mid \delta_n(u) \mid du < \infty$, for all n,

(D2) $\int \delta_n(u)du = 1$,

(D3) $\delta_n(u) \to 0$, uniformly on $\mid u \mid > \eta, \eta > 0$ as $n \to \infty$,

(D4) $\int_{|u|>\eta} \delta_n(u)du \to 0$, for each $\eta > 0$ as $n \to \infty$.

Our choice of kernel $\delta_n(u) = \frac{1}{h^d}K(u/h)$, where h is determined by n as in (A6) can be regarded as a DFS of kernel type.

The following theorem not only provides an exponential-type concentration inequality for $\hat{\mu_1}$ when α is appropriately chosen but also offers a non-asymptotic Bahadur representation result (Bahadur, 1966).

In the theoretical results and their proofs, we will use the notation C to denote positive constant term. Note that the constants $C's$ appearing in different contexts are not necessarily equal.

Theorem 3.3.1. *When conditions (A1) through (A6) are satisfied, let*

$$l(\mu) = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} \psi_\alpha(Y_\mathrm{i} - \mu)\frac{T_\mathrm{i}}{\boldsymbol{\pi}(X_\mathrm{i})}K_h(X_{\mathrm{i}1} - x_1),$$

we have for any $\gamma > 0$, the robust estimator $\hat{\mu}_1(x_1)$ with

$$\alpha = \sqrt{\frac{nE[(Y^{(1)} - \mu_1(x_1))^2 \mid X_1 = x_1]^2}{\gamma C_{K,n}(2)E\left[\frac{(Y^{(1)}-\mu_1(x_1))^2}{\boldsymbol{\pi}(X)} \mid X_1 = x_1\right]}},$$

satisfies the following inequalities:

$$P\left[\mid \hat{\mu}_1(x_1) - \mu_1(x_1) \mid \geq C_1\sqrt{\frac{\gamma}{nh^d}}\right] \leq 6\exp(-\gamma), \tag{3.1}$$

$$P\left[\left|\frac{l(\hat{\mu}_1(x_1))-l(\mu_1(x_1))}{g(x_1)}-(\hat{\mu}_1(x_1)-\mu_1(x_1))\right|\geq C_2\frac{\gamma}{nh^d}\right]\leq 9\exp(-\gamma), \tag{3.2}$$

as long as $n\geq\gamma C_3$*, for some postive constant* C_1, C_2, C_3 *not related to* $\alpha, n,$ *and* γ*.*

An important insight from inequality (3.1) of Theorem 3.3.1 is that even for heavy-tailed errors with only a finite second moment as stated in (A4), the robust estimator $\hat{\mu}_1$ with a properly chosen α exhibits sub-Gaussian tails. As the deviation goes negligible as $n\to\infty$, we conlcude that $\hat{\mu}_1(x_1)$ is consistent.

Furthermore, as shown in (3.2), the remainder term in the Bahadur representation of $\hat{\mu}_1(x_1)$ exhibits sub-exponential tails. As a result, the difference $(\hat{\mu}_1(x_1)-\mu_1(x_1))$ is exponentially close to $\frac{l(\hat{\mu}_1(x_1))-l(\mu_1(x_1))}{g(x_1)}$ with high probability. While the concentration inequality (3.1) provided in Theorem 3.3.1 underscores the effectiveness of our proposed estimator, it may be overly conservative for inference. This is because Bernstein's inequality works best for bounded variables, whereas in our IPW method, the weight $\frac{T_\mathrm{i}}{\boldsymbol{\pi}(X_\mathrm{i})}K_h(X_{\mathrm{i}1}-x_1)$ can become very large when $\boldsymbol{\pi}(X_\mathrm{i})$ is close to zero. Therefore, (3.2) offers a more practical approach to inferenceby approximating the error distribution of the estimator with the linear terms of the Bahadur representation, which are generally well-approximated by a normal distribution.

Since the linear term $\frac{l(\hat{\mu}_1(x_1))-l(\mu_1(x_1))}{g(x_1)}$ can be approximated by a normal distribution when the third moments exist, we are able to derive the Berry-Esseen type bound for $(\hat{\mu}_1-\mu_1(x_1))$ shown in the next theorem, which quantifies the error of the normal approximation in the finite sample case.

Theorem 3.3.2. *Define*

$$\sigma^2(x_1)=Var\left((Y-\mu_1(x_1))\frac{T}{\boldsymbol{\pi}(X)}K_h(X_1-x_1)\right)/g(x_1)^2,$$

$$\sigma_\alpha^2(x_1)=Var\left(\psi_\alpha(Y-\mu_1(x_1))\frac{T}{\boldsymbol{\pi}(X)}K_h(X_1-x_1)\right)/g(x_1)^2.$$

When the conditions in Theorem 3.3.1 are met and the third moment $E[\mid Y^{(1)} - \mu_1(x_1) \mid^3 \mid X_1 = x_1] < \infty$, *then with the same* γ *in Theorem 3.3.1, the statistic*

$$T = \frac{\sqrt{n}(\hat{\mu}_1 - \mu_1(x_1))}{\sigma(x_1)},$$

satisfies

$$\mid P(T \leq t) - \Phi(t) \mid \leq C \left\{ \frac{\gamma}{\sqrt{n}h} + \mathrm{e}^{-\gamma} \right\},$$

for some constant $C > 0$ *not related to* α, n, h, γ.

And the statistic

$$T_1 = \frac{\sqrt{n}(\hat{\mu}_1 - \mu_1(x_1))}{\sigma_\alpha(x_1)},$$

satisfies

$$\mid P(T_1 \leq t) - \Phi(t) \mid \leq C \left\{ \frac{\gamma}{\sqrt{n}h} + \mathrm{e}^{-\gamma} \right\},$$

for some constant $C > 0$ *not related to* α, n, h, γ.

The Berry-Esseen type bound specifies the rate at which the distribution of $\hat{\mu}_1(x_1)$ converges to a normal distribution with mean $\mu_1(x_1)$ and standard deviation $\sigma(x_1)$ or $\sigma_\alpha(x_1)$ as the sample size n increases. The bound is proportional to $\frac{\gamma}{\sqrt{n}h} + \mathrm{e}^{-\gamma}$, implying that for convergence to occur, we require $\gamma = o(\sqrt{n}h)$ and $\gamma \to \infty$.

3.4 Variance Estimator and Confidence Interval

In this section, we focus on inference based on confidence intervals constructed from our proposed estimator. We begin by providing a confidence interval for $\hat{\mu}_1(x_1)$, along with an estimate of its variance. Next, we present a Berry-Esseen-type bound for the CATE estimator $\hat{\tau}(x_1)$, and based on this, we construct a confidence interval for our proposed CATE estimator. The bandwidths used in this section do not necessarily have to be the same as the bandwidths used for the estimation of $\mu_1(x_1)$ and $\mu_0(x_1)$. They should be selected independently in applications.

3.4.1 Confidence Interval for $\mu_1(x_1)$

By Theorem 3.3.2, when α is determined, the variance of the estimator is close to $\sigma^2(x_1)$ and $\sigma_\alpha^2(x_1)$ as $n \to \infty$, where $\sigma^2(x_1)$ reflects the asymptotic case while $\sigma_\alpha^2(x_1)$ as in the finite case with robustification parameter α. Hence to construct a confidence interval, we use an estimator of $\sigma_\alpha^2(x_1)$:

$$\hat{\sigma}_\alpha^2(x_1) = \frac{C_{K,n}(2)\frac{1}{n}\sum_{\mathrm{i}=1}^n \psi_\alpha^2(Y_\mathrm{i} - \hat{\mu}_1(x_1))\frac{T}{\hat{\pi}(X_\mathrm{i})^2}K_h(X_{\mathrm{i}1} - x_1)}{[\frac{1}{n}\sum_{\mathrm{i}=1}^n K_h(X_{\mathrm{i}1} - x_1)]^2}, \tag{3.3}$$

Then based on the Berry-Esseen bound, for significant level $100(1-\beta)\%$, robust confidence intervals can be constructed as:

$$\left[\hat{\mu}_1(x_1) - z_{\beta/2}\frac{\hat{\sigma}_\alpha(x_1)}{\sqrt{n}}, \hat{\mu}_1(x_1) + z_{\beta/2}\frac{\hat{\sigma}_\alpha(x_1)}{\sqrt{n}}\right],$$

where $z_{\beta/2}$ is the $1-\beta/2$th percent quantile.

3.4.2 Confidence Interval for $\tau(x_1)$

Let α_1 and α_0 be the robustification parameters for the treatment group and control group respectively.

Denote $A_\mathrm{i}(x_1) = \psi_{\alpha_1}(Y_\mathrm{i} - \mu_1(x_1))\frac{T_\mathrm{i}}{\pi(X_\mathrm{i})}K_h(X_{\mathrm{i}1} - x_1)/g(x_1)$, $B_\mathrm{i}(x_1) = \psi_{\alpha_0}(Y_\mathrm{i} - \mu_0(x_1))\frac{1-T_\mathrm{i}}{1-\pi(X_\mathrm{i})}$ $K_h(X_{\mathrm{i}1} - x_1)/g(x_1)$, $m_{\alpha_1}(x_1) = E[A_\mathrm{i}(x_1)]$, $m_{\alpha_0}(x_1) = E[B_\mathrm{i}(x_1)]$.

Let

$$T_\tau = \sqrt{n}(\hat{\tau}(x_1) - \tau(x_1))/\sigma_{\tau,\alpha}(x_1),$$

with

$$\sigma_{\tau,\alpha}^2(x_1) = Var(A_\mathrm{i}(x_1) - B_\mathrm{i}(x_1)).$$

Theorem 3.4.1. *Assume the conditions in Theorem 3.3.1 and 3.3.2 are met, with the same γ in Theorem 3.3.1, we have the statistic T_τ satisfies the Berry-Esseen Bound:*

$$sup_t \mid P(T_\tau \le t) - \Phi(t) \mid \le C\left(\frac{\gamma}{\sqrt{n}h} + \mathrm{e}^{-\gamma}\right).$$

Based on Theorem 3.4.1, we're able to construct a $100(1-\beta)\%$ confidence interval for $\hat{\tau}(x_1)$ by :

$$\left[\hat{\tau}(x_1) - z_{\beta/2}\frac{\hat{\sigma}_{\tau,\alpha}(x_1)}{\sqrt{n}}, \hat{\tau}(x_1)(x_1) + z_{\beta/2}\frac{\hat{\sigma}_{\tau,\alpha}(x_1)}{\sqrt{n}}\right],$$

where $z_{\gamma/2}$ is the $\gamma/2$th percent quantile.

By definition, we have $\sigma_{\tau,\alpha}(x_1) = Var(A_{\mathrm{i}}(x_1) - B_{\mathrm{i}}(x_1)) = \sigma^2_{\alpha_1}(x_1) + \sigma^2_{\alpha_1}(x_1) - 2Cov(A_{\mathrm{i}}(x_1), B_{\mathrm{i}}(x_1))$, where $\sigma^2_{\alpha_1}(x_1), \sigma^2_{\alpha_0}(x_1)$ can be estimated by (3.3).

Since $m_{\alpha_1} = O(h_1^s), m_{\alpha_0} = O(h_0^s)$, we have $Cov(A_{\mathrm{i}}(x_1), B_{\mathrm{i}}(x_1)) \approx E[A_{\mathrm{i}}(x_1)B_{\mathrm{i}}(x_1)]$.

Similar to the lemma 2.1 in Härdle (1984), we have $\{\delta_{n,\tau}(u)\} = \{\mid \delta_{n,1}(u)\delta_{n,2}(u) \mid /C_{K,\tau}\}$ is also a DFS, where $C_{K,\tau} = \int \mid \delta_{n,1}(u)\delta_{n,2}(u) \mid du < \infty$. Then we can estimate $Cov(A_{\mathrm{i}}(x_1), B_{\mathrm{i}}(x_1))$ by:

$$\widehat{Cov}(A_{\mathrm{i}}(x_1), B_{\mathrm{i}}(x_1)) = \frac{C_{K,\tau}\frac{1}{n}\sum_{\mathrm{i}=1}^n \hat{A}_{\mathrm{i}}\hat{B}_{\mathrm{i}}K_h(X_{\mathrm{i}1} - x_1)}{\frac{1}{n}\sum_{\mathrm{i}=1}^n K_h(X_{\mathrm{i}1} - x_1))^2},$$

where

$$\begin{aligned} A_{\mathrm{i}}(x_1) &= \psi_{\alpha_1}(Y_{\mathrm{i}} - \hat{\mu}_1(x_1))\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}K_h(X_{\mathrm{i}1} - x_1)/g(x_1), B_{\mathrm{i}}(x_1) \\ &= \psi_{\alpha_0}(Y_{\mathrm{i}} - \hat{\mu}_0(x_1))\frac{1 - T_{\mathrm{i}}}{1 - \boldsymbol{\pi}(X_{\mathrm{i}})}K_h(X_{\mathrm{i}1} - x_1)/g(x_1). \end{aligned}$$

3.5 Algorithm

To provide a point estimator with a valid confidence interval, we select the robustification parameter α based on our theoretical conditions. According to Theorem 3.3.1, we set

$$\alpha = \sqrt{\frac{nE[(Y^{(1)} - \mu_1(x_1))^2 \mid X_1 = x_1]g(x_1)}{\gamma C_{K,n}(2)E[(Y^{(1)-\mu_1(x_1)})^2/\boldsymbol{\pi}(X) \mid X_1 = x_1]}},$$

and according to Theorem 3.3.2, we take $\gamma = n^{2/5}h = o(\sqrt{n}h)$. Since $\hat{\alpha}$ and $\hat{\mu}_1(x_1)$ are interdependent, $\hat{\alpha}$ is iteratively selected in each loop of an IRLS procedure, as shown below.

Begin with an initial estimator $\hat{\mu}_1^{(0)}$ from either the L_2 loss or the L_1 loss, we can estimate the robustification parameter and then estimate the μ_1.

Let $\hat{\mu}_1^{(t)}(x_1)$ be the solution from the t-th iteration, in the $t+1$-th loop, $\alpha^{(t+1)}$ can be estimated by:

$$\hat{\alpha}^{(t+1)} = \sqrt{\frac{nE_1^{(t+1)}\hat{g}(x_1)}{\gamma C_{K,n}(2)E_2^{(t+1)}}},$$

where

$$E_1^{(t+1)} = \frac{1}{nh}\sum_{\mathrm{i}=1}^{n}(Y_\mathrm{i} - \hat{\mu}_1^{(t)})^2\frac{T_\mathrm{i}}{\hat{\boldsymbol{\pi}}(X_\mathrm{i})}K_h(X_{\mathrm{i}1} - x_1),$$

$$E_2^{(t+1)} = \frac{1}{nh}\sum_{\mathrm{i}=1}^{n}(Y_\mathrm{i} - \hat{\mu}_1^{(t)})^2\frac{T_\mathrm{i}}{\hat{\boldsymbol{\pi}}(X_\mathrm{i})^2}K_h(X_{\mathrm{i}1} - x_1),$$

and

$$\hat{g}(x_1) = \frac{1}{nh}\sum_{\mathrm{i}=1}^{n}K_h(X_{\mathrm{i}1} - x_1),$$

and then $\hat{\mu}_1^{(t+1)}$ can be solved by the following estimating:

$$0 = \sum_{\mathrm{i}=1}^{n}\frac{T_\mathrm{i}K_h(X_{\mathrm{i}1} - x_1)}{\hat{\boldsymbol{\pi}}(X_\mathrm{i})}W(Y_\mathrm{i} - \hat{\mu}_1^{(t)}(x_1))(Y_\mathrm{i} - \mu_1^{(t+1)}(x_1)),$$

where $W(x) = \begin{cases} \psi_\alpha(x)/x & x \neq 0 \\ \psi'_\alpha(x) & x = 0 \end{cases}$.

The conditional mean of other potential outcome $\mu_0(x_1)$ can be estimated similarly to $\mu_1(x_1)$. Once $\hat{\mu}_1(x_1)$ and $\hat{\mu}_0(x_1)$ are estimated, we can then estimate $\hat{\tau}(x_1) = \hat{\mu}_1(x_1) - \hat{\mu}_0(x_1)$.

3.6 Simulation

In this section, we perform Monte Carlo simulations to evaluate the properties of the estimator $\hat{\tau}(x_1)(x_1)$. We compare its performance with the regular Huber loss and the L_2 loss, as studied in Chapter 2, to demonstrate and verify the behavior of our proposed estimator, which balances the properties of both L_1 and L_2 losses. Since a comprehensive comparison between L_1, L_2, and other loss functions, along with methods like IPW and

AIPW, has already been provided in Chapter 2, we omit such comparisons here, as the results are predictable based on the earlier work.

3.6.1 Data Generating Process

The data in this study is generated using a simple causal setting. The covariates (X_1, X_2, X_3, X_4) are independently drawn from uniform distributions: $X_{\mathrm{i}1}, X_{2\mathrm{i}}, X_{3\mathrm{i}}, X_{4\mathrm{i}} \overset{\text{i.i.d.}}{\sim}$ Unif$(-0.5, 0.5)$. The treatment indicator T_{i} is generated as $T_{\mathrm{i}} \sim$ Bernoulli$(\Lambda(0.5(X_{\mathrm{i}1} + X_{2\mathrm{i}} + X_{3\mathrm{i}} + X_{4\mathrm{i}})))$, where Λ denotes the logistic function. The observed outcome Y is generated according to $Y_{\mathrm{i}} = Y_{\mathrm{i}}^{(1)} D_{\mathrm{i}} + Y_{\mathrm{i}}^{(0)}(1 - D_{\mathrm{i}})$, with potential outcomes $(Y^{(1)}, Y^{(0)})$ generated by $Y_{\mathrm{i}}^{(1)} = 10 - 5X_{\mathrm{i}1} - \cos(5X_{\mathrm{i}1}) - 5X_{2\mathrm{i}} - 3\tanh(0.5X_{3\mathrm{i}}) + \epsilon_{\mathrm{i}}$, and $Y_{\mathrm{i}}^{(0)} = 0$.

The error term ϵ_{i} are generated from various distributions, representing different types of tail behavior and symmetry property:

(a)Symmetric light-tail: Normal errors with mean 0 and variance 10 $N(0, 10)$;

(b)Symmetric heavy-tail: Scaled t-distribution errors $(2 \cdot t_3)$;

(c)Asymmetric heavy-tail: Log-normal errors $(\log N(1, 1.2) - \exp(1 + 1.2^2/2))$ and Weibull errors $(\text{Weibull}(0.3, 0.5) - 0.5\Gamma\left(1 + \frac{1}{0.3}\right))$;

(d)Asymmetric light-tail: Mixed normal distribution $(D \cdot N(-1, 2) + (1 - D) \cdot N(5, 1) - 2$ with $D \sim \text{Bern}(0.5))$.

3.6.2 Simulation Method

To verify the theoretical properties of $\hat{\mu}_1(x_1)$, we estimate $\mu_1(x_1)$ for x_1 taken values from $\{-0.4, -0.2, 0, 0.2, 0.4\}$ using sample sizes of $n = 500$ and $n = 5000$. The number of Monte Carlo replications is 2000. Motivated by Theorem 3.3.2, we choose $x = n^{2/5}h$ for the rest of the cases to ensure asymptotic normality. For all estimators and variance estimates, we use normal kernels with bandwidth h selected by cross-validation.

The simulation results for each sample size are summarized in two tables. For each x_1, the first table includes the true value, the mean of the estimated $\hat{\tau}(x_1)(x_1)$, the mean of the estimated variance based on Theorem 3.3.2, the bias, variance, and mean squared error

(MSE) of our proposed estimator. Additionally, these results are compared with the fixed Huber loss and the L_2 loss by estimating their bias, variance, and MSE.

Confidence intervals are constructed based on the conditional mean estimates and their corresponding estimated variances. For each method, the coverage rate and the mean width of the confidence intervals are reported.

As expected, the bias and variance of $\hat{\mu}_1(x_1)$ is greater when x_1 is close to the boundary, hence we will mainly focus on the estimators on $x_1 \in \{-0.2, 0, 0.2\}$.

Across different sample sizes, there is a notable improvement in both the bias and variance of our proposed method across all error distributions. The validity of the variance estimator is confirmed, as the mean of the estimated variances closely matches the simulation variance. Compared to other methods, our proposed method outperforms the L_2 loss in all scenarios involving heavy-tailed errors. In light-tailed cases, the performance of the three methods is comparable. In heavy-tailed and asymmetric cases, our proposed estimator outperforms the fixed Huber loss due to its effectiveness in handling the unshrinkable biases present in the fixed Huber loss.

With our choice of bandwidth and robustification parameter, the coverage rate of the confidence interval for our proposed method is reasonably close to the true values when $n = 5000$, and the confidence intervals provide reliable inference without major coverage rate distortions.

All methods perform similarly under light-tailed errors. Our proposed method performs similar to the fixed Huber estimator under heavy-tailed and symmetric errors, with mean confidence interval widths smaller than the L_2 loss but larger than the Huber loss. With heavy-tailed and asymmetric errors, the coverage rate of the fixed Huber loss estimators decreases as the sample size increases due to constant bias, while our proposed estimator's coverage rate increases and approaches the expected rate as n goes to infinity. For the log-normal distribution, the coverage rate is close to 90% and 95% for $n = 5000$, and for the Weibull distribution, similar results can be achieved by increasing the sample size or by reducing the value of γ.

From the second table, all confidence intervals provide valid inferences with light-tailed error distributions. In heavy-tailed cases, the estimators with Huber loss and our proposed

Table 3.1. Bias, Variance, and Mean Squared Error (MSE) for $\tau_1(x_1)$ Estimates with Sample Size $n = 500$.

	Adpt				Huber			L2		
x_1	Bias	Variance	EstVar	MSE	Bias	Variance	MSE	Bias	Variance	MSE
Log-Normal										
-0.4	-0.924	0.931	0.726	1.785	-1.548	0.452	2.849	-0.957	1.001	1.916
-0.2	-0.144	0.603	0.503	0.624	-1.201	0.328	1.770	0.006	0.718	0.718
0.0	0.113	0.534	0.460	0.547	-1.125	0.307	1.574	0.413	0.638	0.808
0.2	0.013	0.542	0.461	0.542	-1.193	0.289	1.711	0.291	0.628	0.713
0.4	-0.113	0.767	0.619	0.779	-1.265	0.336	1.936	0.079	0.785	0.791
Weibull										
-0.4	-1.509	3.580	2.496	5.856	-3.175	0.277	10.355	-0.932	6.038	6.906
-0.2	-0.693	2.303	1.586	2.783	-3.032	0.235	9.428	-0.031	3.974	3.975
0.0	-0.385	1.929	1.387	2.077	-3.001	0.217	9.224	0.311	3.528	3.625
0.2	-0.470	2.074	1.491	2.296	-3.043	0.190	9.449	0.216	3.975	4.021
0.4	-0.567	2.982	2.174	3.304	-3.069	0.223	9.640	0.071	5.860	5.865
Mixture of Normals										
-0.4	-0.440	0.204	0.191	0.398	-0.503	0.191	0.444	-0.437	0.192	0.383
-0.2	0.064	0.147	0.151	0.152	0.062	0.123	0.127	0.060	0.135	0.139
0.0	0.158	0.119	0.136	0.144	0.188	0.112	0.148	0.163	0.125	0.152
0.2	0.098	0.122	0.127	0.132	0.110	0.113	0.125	0.093	0.127	0.136
0.4	0.002	0.172	0.151	0.172	0.015	0.149	0.150	-0.005	0.157	0.157
T-distribution										
-0.4	-0.631	0.506	0.503	0.904	-0.657	0.394	0.826	-0.608	0.644	1.015
-0.2	0.070	0.347	0.369	0.352	0.075	0.259	0.265	0.083	0.413	0.420
0.0	0.233	0.327	0.336	0.381	0.280	0.237	0.316	0.278	0.378	0.455
0.2	0.130	0.344	0.323	0.361	0.166	0.229	0.256	0.168	0.387	0.415
0.4	0.016	0.462	0.411	0.463	0.011	0.319	0.319	-0.003	0.506	0.506
Normal										
-0.4	-0.416	0.187	0.171	0.360	-0.416	0.185	0.358	-0.365	0.191	0.324
-0.2	0.078	0.121	0.135	0.127	0.077	0.128	0.134	0.068	0.134	0.138
0.0	0.162	0.108	0.122	0.135	0.160	0.121	0.147	0.142	0.121	0.141
0.2	0.089	0.109	0.113	0.117	0.087	0.111	0.118	0.074	0.119	0.124
0.4	-0.001	0.148	0.135	0.148	-0.002	0.141	0.141	-0.006	0.161	0.161

Table 3.2. Confidence Interval Coverage and Width for $\tau_1(x_1)$ with Sample Size $n = 500$.

	Adpt				Huber				L2			
x_1	Rate90	width90	Rate95	width95	Rate90	width90	Rate95	width95	Rate90	width90	Rate95	width95
logN												
-0.4	0.565	2.650	0.639	3.157	0.224	1.957	0.305	2.332	0.580	2.873	0.654	3.424
-0.2	0.814	2.257	0.875	2.689	0.292	1.746	0.387	2.081	0.850	2.513	0.909	2.994
0.0	0.868	2.161	0.925	2.575	0.318	1.683	0.409	2.006	0.885	2.417	0.951	2.880
0.2	0.851	2.159	0.913	2.573	0.270	1.641	0.348	1.955	0.891	2.409	0.946	2.871
0.4	0.820	2.470	0.879	2.943	0.273	1.753	0.348	2.089	0.859	2.691	0.915	3.207
Weibull												
-0.4	0.475	4.258	0.522	5.073	0.002	1.587	0.005	1.891	0.604	5.890	0.649	7.019
-0.2	0.620	3.624	0.676	4.318	0.000	1.484	0.002	1.768	0.775	5.254	0.814	6.261
0.0	0.692	3.480	0.760	4.147	0.001	1.445	0.002	1.721	0.843	5.083	0.892	6.057
0.2	0.669	3.533	0.726	4.210	0.000	1.385	0.000	1.650	0.799	5.075	0.853	6.047
0.4	0.647	4.101	0.701	4.887	0.002	1.417	0.002	1.688	0.745	5.660	0.796	6.744
MixN												
-0.4	0.734	1.433	0.814	1.708	0.671	1.405	0.785	1.674	0.734	1.438	0.829	1.713
-0.2	0.894	1.274	0.946	1.518	0.916	1.237	0.956	1.474	0.910	1.272	0.955	1.516
0.0	0.881	1.212	0.933	1.445	0.874	1.169	0.935	1.393	0.872	1.212	0.930	1.444
0.2	0.882	1.171	0.942	1.395	0.890	1.128	0.933	1.344	0.880	1.170	0.938	1.394
0.4	0.875	1.274	0.930	1.519	0.896	1.241	0.942	1.479	0.888	1.279	0.944	1.524
T3												
-0.4	0.750	2.293	0.829	2.732	0.686	1.984	0.792	2.364	0.740	2.390	0.833	2.848
-0.2	0.910	1.974	0.961	2.352	0.894	1.706	0.952	2.033	0.904	2.073	0.947	2.470
0.0	0.875	1.886	0.932	2.247	0.840	1.619	0.909	1.929	0.858	1.996	0.922	2.378
0.2	0.883	1.846	0.939	2.199	0.877	1.588	0.935	1.892	0.883	1.950	0.937	2.324
0.4	0.876	2.075	0.935	2.473	0.886	1.790	0.941	2.133	0.881	2.165	0.939	2.580
N												
-0.4	0.712	1.353	0.806	1.612	0.721	1.368	0.803	1.630	0.765	1.395	0.849	1.663
-0.2	0.905	1.205	0.944	1.436	0.903	1.213	0.948	1.445	0.908	1.251	0.952	1.491
0.0	0.874	1.146	0.936	1.365	0.861	1.148	0.927	1.368	0.882	1.192	0.936	1.421
0.2	0.890	1.102	0.944	1.313	0.886	1.104	0.947	1.315	0.893	1.151	0.949	1.372
0.4	0.883	1.202	0.941	1.432	0.885	1.204	0.941	1.434	0.877	1.243	0.936	1.482

Table 3.3. Bias, Variance, and Mean Squared Error (MSE) for $\tau_1(x_1)$ Estimates with Sample Size $n = 5000$.

	Adpt				Huber			L2		
x_1	Bias	Variance	EstVar	MSE	Bias	Variance	MSE	Bias	Variance	MSE
Log-Normal										
-0.4	-0.481	0.132	0.108	0.363	-1.382	0.050	1.960	-0.356	0.154	0.281
-0.2	0.002	0.091	0.083	0.091	-1.233	0.045	1.566	0.079	0.113	0.119
0.0	0.100	0.077	0.078	0.087	-1.209	0.040	1.501	0.139	0.116	0.135
0.2	0.034	0.077	0.075	0.078	-1.236	0.036	1.564	0.073	0.101	0.107
0.4	-0.054	0.107	0.091	0.110	-1.280	0.041	1.680	-0.002	0.124	0.124
Weibull										
-0.4	-0.970	0.503	0.366	1.443	-3.097	0.040	9.633	-0.807	0.683	1.333
-0.2	-0.226	0.329	0.262	0.381	-3.080	0.037	9.523	0.037	0.441	0.443
0.0	0.003	0.303	0.244	0.303	-3.065	0.035	9.430	0.344	0.396	0.514
0.2	-0.109	0.311	0.242	0.323	-3.071	0.031	9.461	0.212	0.418	0.463
0.4	-0.249	0.424	0.316	0.486	-3.087	0.031	9.558	0.029	0.550	0.551
Mixture of Normals										
-0.4	-0.069	0.031	0.031	0.036	-0.121	0.029	0.044	-0.128	0.029	0.046
-0.2	0.021	0.027	0.028	0.027	0.032	0.024	0.025	0.026	0.024	0.025
0.0	0.041	0.027	0.026	0.028	0.059	0.022	0.026	0.060	0.023	0.026
0.2	0.025	0.024	0.025	0.024	0.033	0.022	0.023	0.027	0.021	0.021
0.4	-0.009	0.025	0.025	0.025	-0.015	0.024	0.024	-0.012	0.023	0.024
T-distribution										
-0.4	-0.237	0.081	0.075	0.137	-0.166	0.063	0.091	-0.221	0.091	0.140
-0.2	0.053	0.060	0.062	0.063	0.043	0.048	0.049	0.047	0.068	0.070
0.0	0.095	0.055	0.058	0.064	0.079	0.046	0.053	0.085	0.062	0.069
0.2	0.060	0.053	0.055	0.057	0.037	0.045	0.046	0.042	0.055	0.057
0.4	-0.002	0.068	0.061	0.068	-0.009	0.050	0.050	-0.018	0.070	0.070
Normal										
-0.4	-0.115	0.024	0.025	0.038	-0.119	0.026	0.041	-0.124	0.026	0.041
-0.2	0.030	0.021	0.022	0.022	0.034	0.022	0.023	0.032	0.021	0.022
0.0	0.060	0.020	0.020	0.024	0.059	0.021	0.024	0.056	0.019	0.022
0.2	0.031	0.020	0.019	0.021	0.031	0.020	0.020	0.033	0.019	0.021
0.4	-0.016	0.021	0.020	0.022	-0.008	0.020	0.020	-0.008	0.022	0.022

Table 3.4. Confidence Interval Coverage and Width for $\tau_1(x_1)$ with Sample Size $n = 500$.

	Adpt				Huber				L2			
x_1	Rate90	width90	Rate95	width95	Rate90	width90	Rate95	width95	Rate90	width90	Rate95	width95
logN												
-0.4	0.520	1.072	0.614	1.278	0.000	0.699	0.000	0.833	0.667	1.213	0.751	1.446
-0.2	0.883	0.942	0.937	1.122	0.000	0.657	0.000	0.783	0.912	1.093	0.957	1.303
0.0	0.894	0.913	0.951	1.088	0.000	0.636	0.000	0.758	0.885	1.059	0.945	1.262
0.2	0.900	0.894	0.943	1.065	0.000	0.619	0.000	0.737	0.899	1.029	0.953	1.226
0.4	0.860	0.982	0.912	1.170	0.000	0.630	0.000	0.750	0.884	1.103	0.932	1.314
Weibull												
-0.4	0.452	1.918	0.524	2.286	0.000	0.629	0.000	0.750	0.555	2.350	0.624	2.801
-0.2	0.784	1.644	0.840	1.959	0.000	0.608	0.000	0.725	0.869	2.032	0.918	2.421
0.0	0.849	1.589	0.902	1.894	0.000	0.592	0.000	0.706	0.886	1.956	0.945	2.331
0.2	0.817	1.582	0.886	1.885	0.000	0.579	0.000	0.690	0.880	1.943	0.938	2.315
0.4	0.752	1.795	0.822	2.139	0.000	0.566	0.000	0.675	0.850	2.174	0.902	2.591
MixN												
-0.4	0.873	0.582	0.935	0.693	0.800	0.548	0.881	0.653	0.795	0.549	0.873	0.654
-0.2	0.909	0.553	0.959	0.659	0.889	0.514	0.946	0.613	0.891	0.513	0.943	0.612
0.0	0.890	0.535	0.939	0.637	0.871	0.496	0.925	0.592	0.874	0.496	0.928	0.591
0.2	0.901	0.520	0.955	0.620	0.884	0.482	0.939	0.574	0.901	0.482	0.954	0.574
0.4	0.898	0.521	0.953	0.621	0.885	0.490	0.939	0.584	0.887	0.490	0.945	0.584
T3												
-0.4	0.759	0.897	0.841	1.069	0.803	0.798	0.875	0.951	0.783	0.931	0.862	1.109
-0.2	0.904	0.817	0.952	0.974	0.899	0.737	0.954	0.878	0.892	0.847	0.950	1.010
0.0	0.880	0.792	0.939	0.944	0.883	0.715	0.945	0.852	0.885	0.816	0.936	0.972
0.2	0.892	0.773	0.947	0.921	0.893	0.696	0.946	0.829	0.910	0.799	0.956	0.952
0.4	0.884	0.813	0.933	0.968	0.895	0.718	0.949	0.856	0.891	0.834	0.938	0.993
N												
-0.4	0.812	0.517	0.879	0.616	0.793	0.518	0.877	0.618	0.790	0.518	0.871	0.617
-0.2	0.903	0.486	0.949	0.579	0.895	0.487	0.948	0.580	0.907	0.486	0.952	0.579
0.0	0.880	0.469	0.928	0.558	0.865	0.469	0.921	0.559	0.889	0.468	0.948	0.558
0.2	0.882	0.454	0.942	0.541	0.892	0.455	0.946	0.542	0.886	0.454	0.941	0.541
0.4	0.880	0.463	0.940	0.552	0.900	0.463	0.951	0.552	0.878	0.463	0.938	0.552

method yield narrower confidence intervals compared to the L_2 loss. With heavy-tailed and asymmetric error distributions, the coverage rate of the Huber loss decreases as the sample size increases, while the confidence intervals based on our proposed estimator, though slightly wider, achieve coverage rates approaching 90% and 95% as the sample size increases.

In conclusion, the simulation results indicate that between the unbiased L2 estimator and the robust Huber loss, the adaptive Huber loss estimator achieves bias reduction similar to the L_2 loss and maintains a certain level of robustness akin to the Huber loss when the robustification parameter is appropriately chosen.

3.7 Data Application

In this section, we examine the CATE of alcohol consumption on liver function biomarkers across different age groups. The primary biomarkers considered are alanine aminotransferase (ALT), aspartate aminotransferase (AST), gamma-glutamyl transferase (GGT), and alkaline phosphatase (ALP). Previous research indicates that alcohol drinking is associated with elevated levels of serum GGT, ALT, and AST, and ALP, particularly among heavy drinkers (Torkadi, Apte, and Bhute, 2014; Lala, Zubair, and Minter, 2023; Conigrave et al., 2003; Agarwal, Fulgoni, and Lieberman, 2015). It has also been noted that the impact of alcohol on liver enzymes is more pronounced in older adults and those with a longer history of alcohol use (Alatalo et al., 2009; Hietala et al., 2005). Furthermore, Agarwal, Fulgoni, and Lieberman (2015) demonstrated significant effects of age on liver enzymes. Despite this, there is limited research investigating the heterogeneity of the effect of alcohol on liver function across various age groups. Using CATE allows for a more nuanced understanding of this age-related heterogeneity compared to simple sample splitting. To address this gap and to illustrate the application of our proposed method, we conduct four separate estimations to assess the CATE of alcohol consumption on each of the four liver function markers (ALT, AST, GGT, ALP), conditioned on age.

3.7.1 Dataset and Estimation Strategy

To estimate the CATE of self-reported alcohol consumption on liver function markers, we utilize data from the NHANES for adults, spanning 1999-2006. This dataset provides sufficient information to perform our analysis and benefits from being collected using consistent, standardized procedures over multiple years.

For the treatment indicator, we used the criteria set by the CDC: heavy drinking is defined as consuming eight or more drinks per week for women or 15 or more drinks per week for men (Stahre et al., 2006; C. o. State et al., 2004; National Institute of Alcohol Abuse and Alcoholism (NIAAA), 2004; Agriculture, Health, and Services, 2005; Health and Services, 2000). Accordingly, we categorized individuals as treatment if they reported consuming more than two alcoholic drinks per day in the past 12 months for males, more than one alcoholic drink per day for females, and as control if they reported fewer drinks per day.

Given the large number of variables and their varying patterns of missingness, including all variables would significantly reduce the sample. Therefore, we selected variables routinely used in the literature (Agarwal, Fulgoni, and Lieberman, 2015; Åberg, Färkkilä, and Männistö, 2020): race/ethnicity, age, physical activity (categorized as sedentary, moderate, or vigorous), poverty income ratio, smoking habits (yes/no), energy intake (kcal), and BMI. We assume unconfoundedness with respect to these covariates, as indicated in the references. After excluding missing values, the final sample size is 290 in the treatment group and $1,661$ in the control group.

To gain a general understanding of the data, we create scatter plots, histograms, and Q-Q plots for each of the responses. As commonly observed in each of the scatter plots, some data points deviate noticeably from the majority towards the positive side, clearly indicating the presence of asymmetrically distributed outliers. The histograms and Q-Q plots further confirm that the distribution of the responses exhibits pronounced tails and asymmetry. These findings suggest both heavy-tailedness and skewness, underscoring the necessity of our proposed method to account for these characteristics.

The age range in the dataset spans from 20 to 80 years. To mitigate boundary effects, we estimate the CATE within a grid of ages from 25 to 75. We estimate the CATE for the four

liver function markers at 50 equally spaced grid points within this interval. The propensity score is estimated using logistic regression. For the estimation of μ_1 and μ_0, we employ a regular Gaussian kernel, and the bandwidths are selected via cross-validation.

The point estimates are presented in the plots with red lines, and the black dashed lines represent the estimated 95% confidence intervals.

3.7.2 Estimation Results

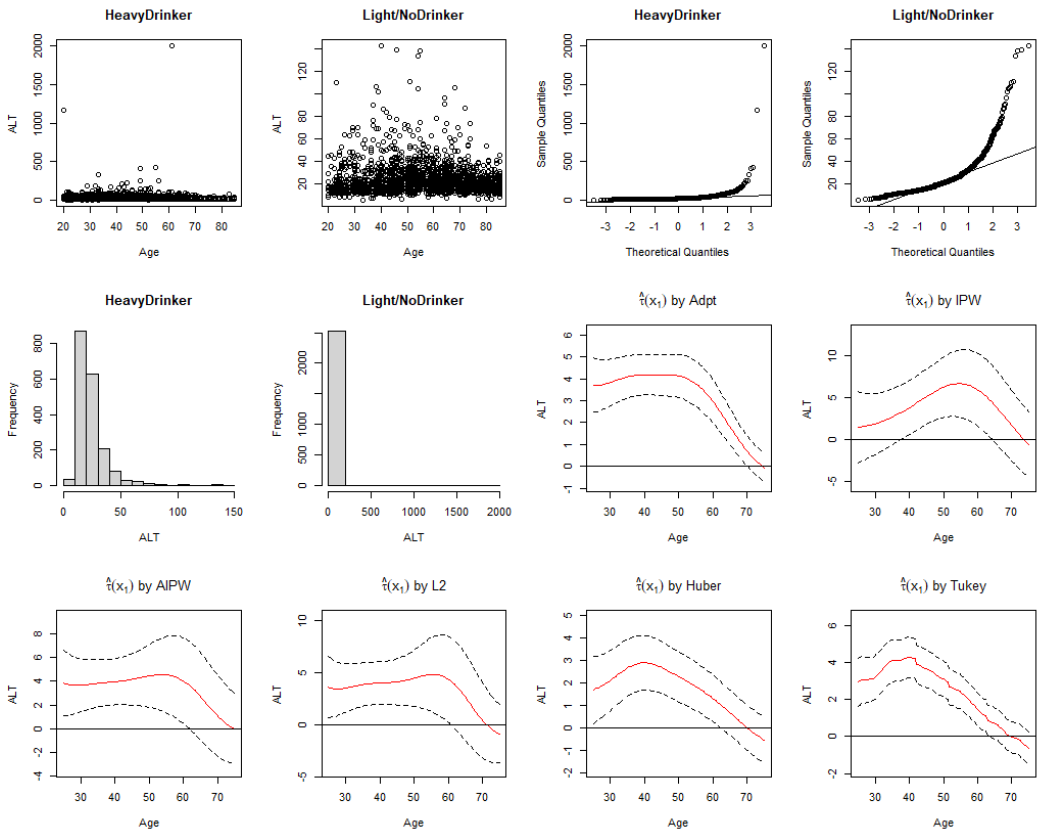

Figure 3.1. Scatter Plot, QQ-plot, histogram and CATE estiamtors with 95% confidence interval for ALT.

The CATE estimator for ALT reveals a slight increase before age 40 and begins to decrease after age 50. It remains positive for individuals of all ages except after age 74. The confidence interval indicates that the effect is significantly positive for individuals younger than 70 but not significant for older age groups. This is consistent with existing studies showing that

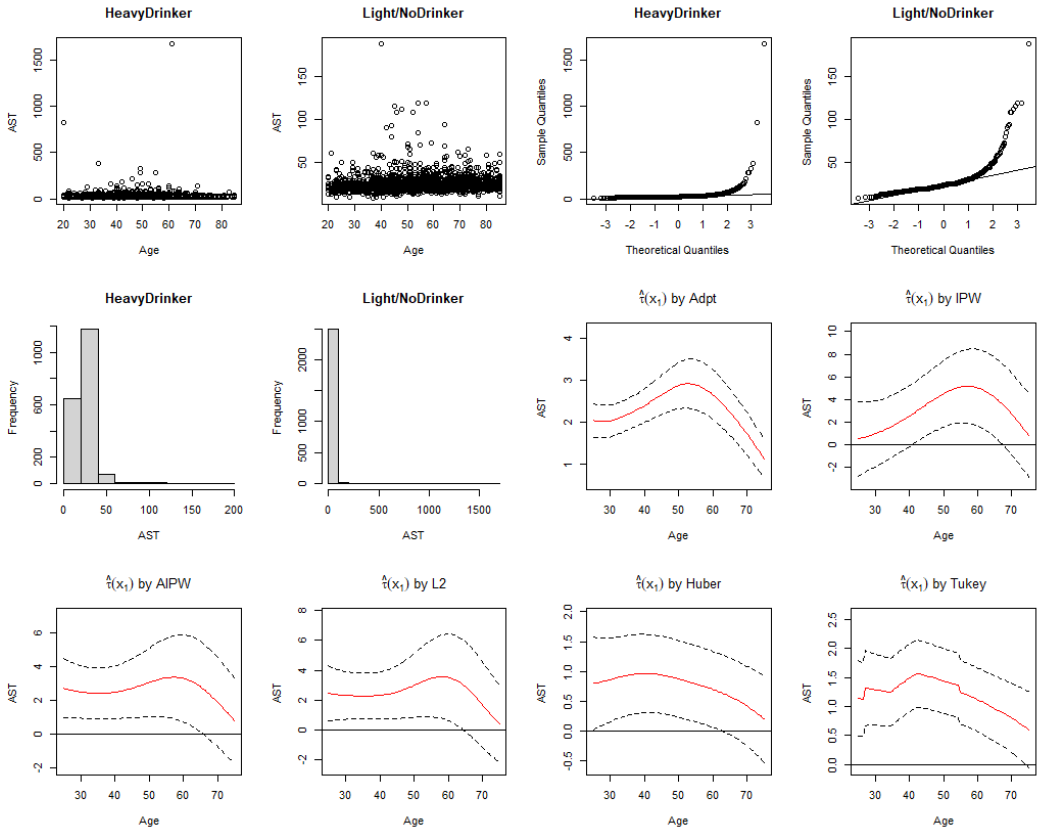

Figure 3.2. Scatter Plot, QQ-plot, histogram and CATE estiamtors with 95% confidence interval for AST.

ALT levels typically rise with alcohol consumption; however, this increase can diminish when liver cells are damaged, as discussed by Giannini et al. (1999). The CATE estimator for AST shows a positive effect across all ages, with the confidence interval indicating 95% significance. For GGT, the CATE estimator reveals a slight increase before age 45 and a decrease thereafter. The confidence interval suggests 95% significance for all age groups except those older than 72. The CATE estimator for ALP demonstrates a positive effect that increases before age 40 and decreases thereafter, with the confidence interval indicating 95% significance from ages 30 to 56.

When comparing the results from different methods, all estimators exhibit similar trends and patterns. However, the IPW method is particularly sensitive to the choice of bandwidth, often producing different estimated values with varying bandwidth selections. The AIPW

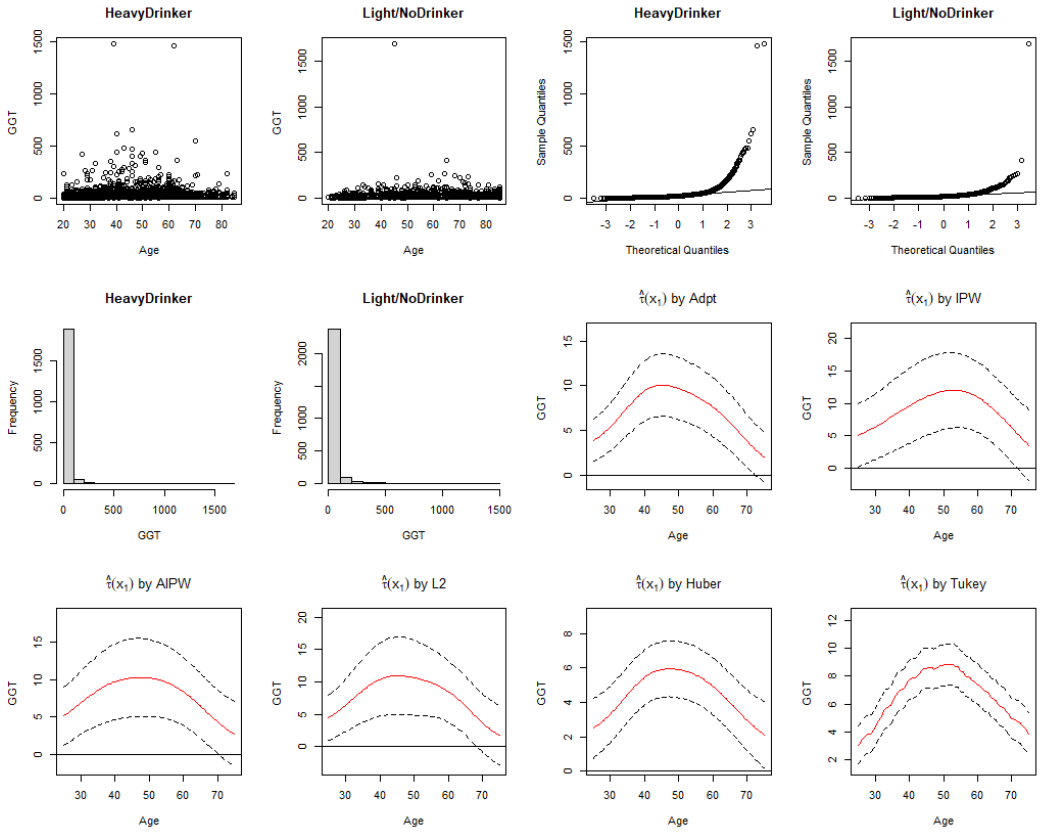

Figure 3.3. Scatter Plot, QQ-plot, histogram and CATE estiamtors with 95% confidence interval for GGT.

estimator, along with the method we proposed in study 1 using the L_2 loss, shows similar trends and magnitudes to our proposed method with adaptive Huber loss, but with wider confidence intervals. In comparison to robust estimators with fixed robust loss functions, we observe that although the confidence intervals are narrower, these fixed-loss robust estimators tend to yield smaller values than both our proposed estimator and the non-robust estimators. This suggests the presence of nonreducible bias resulting from the fixed loss functions under asymmetric error distributions.

Overall, the CATE estimators suggest that heavy drinking has positive effects on ALT, AST, GGT, and ALP across nearly all age groups. These results align with existing literature, as the four biomarkers are known to increase with heavy drinking (Lala, Zubair, and Minter, 2023; Agarwal, Fulgoni, and Lieberman, 2015). A common trend observed across the

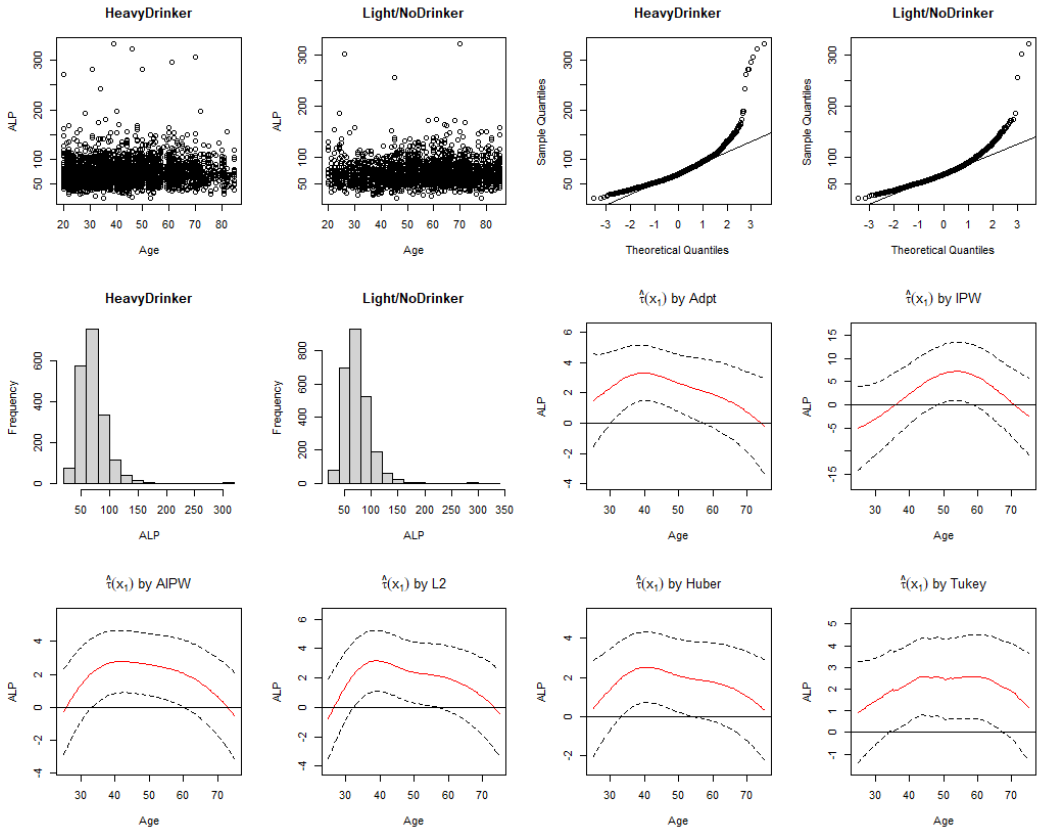

Figure 3.4. Scatter Plot, QQ-plot, histogram and CATE estiamtors with 95% confidence interval for ALP.

four results is that middle-aged individuals tend to have higher levels of ALT, AST, GGT, and ALP compared to both younger and older individuals. A plausible explanation for the increasing trend in younger ages is that long term heavy drinking may have a cumulative negative effect on liver function. For the decreasing trend observed in older individuals, Mowé and Bøhmer (1996) suggests that the increase in alcohol biomarkers may not be directly related to heavy drinking. Another possibility is the presence of unmeasured confounders, where the correlation between heavy drinking and other harmful behaviors associated with age is not fully captured, potentially leading to some selection bias.

3.8 Conclusion

In this paper, we extended the study of CATE estimation to settings involving heavy-tailed error distributions without assuming symmetry. Building on Chapter 2 with local M-estimators and inverse propensity weighting, we applied the adaptive Huber estimator, where the robustification parameter diverges with the sample size to effectively balance bias and robustness. Our key theoretical contributions include the derivation of concentration inequalities for the proposed estimator and for the remainder of the Bahadur representation, based on the inequalities, we further derived a Berry-Esseen type inequality, providing essential insights into its properties and guiding the selection of the robustification parameter for asymptotic unbiasedness and for inference purpose. Furthermore, we derived asymptotic normal-based confidence intervals, enhancing the practical applicability of our method in statistical inference.

Our proposed method achieves both asymptotic unbiasedness and robustness, compared to traditional robust methods based on regular loss functions and non-robust methods based on squared loss. Additionally, since the robustification parameter is directly estimated based on the conditions derived from our theoretical results, the computational time is significantly lower compared to other existing robust methods for handling asymmetric, heavy-tailed error distributions.

Our proposed method provides more effective confidence intervals for asymmetric and heavy-tailed distributions with a bell shape, such as the log-normal distribution. However, for more extreme heavy-tailed and asymmetric distributions (like the Weibull distribution), a larger robustification parameter or a significantly larger sample size may be necessary for the coverage rate of the confidence intervals to approach the nominal level. We plan to investigate solutions to this issue in future research.

4. NONPARAMETRIC OUTLIER RESISTANT CONDITIONAL AVERAGE TREATMENT EFFECT ESTIMATOR WITH SDR

The study focuses on constructing an outlier-resistant estimator for the CATE by integrating the M-estimator approach with the inverse propensity score weighting (IPW) method. Specifically, we propose a fully nonparametric method using the Nadaraya-Watson (NW) estimator with higher-order kernels. To address the limitations of this method in high-dimensional settings, we replace the full set of covariates with lower-dimensional subsets for estimating the conditional mean of potential outcomes. These lower-dimensional subsets are determined using SDR techniques and are outcome-specific. We derive the asymptotic properties of our proposed estimator and conduct an efficiency comparison across estimators that utilize different SDR subspaces.

4.1 Introduction

Estimating the CATE offers valuable insights into treatment effects across diverse subpopulations. Under the counterfactual framework (Rubin, 1974; Neyman, 1923), Abrevaya, Hsu, and Lieli (2015) proposed an inverse propensity weighting (IPW) estimator for situations where the unconfoundedness assumption does not hold generally, conditional on low-dimensional covariates. However, the IPW method proposed by Abrevaya, Hsu, and Lieli (2015) is sensitive to outliers and exhibits large variance when dealing with heavy-tailed error distributions. Consequently, the performance of these methods is limited, necessitating the development of an outlier-resistant approach.

To address this challenge, in Chapter 2, we constructed an outlier-resistant estimator for the CATE conditioned on a low-dimensional subset of covariates, following the semi-parametric method proposed by Abrevaya, Hsu, and Lieli (2015). The CATE was estimated using the Nadaraya-Watson estimator (NW-estimator) (Nadaraya, 1964; Watson, 1964), while the propensity score, treated as a nuisance parameter, was estimated using a parametric

method such as logistic regression. However, relying on a parametric model for propensity score estimation may lead to misleading results if the model is misspecified.

Alternatively, Abrevaya, Hsu, and Lieli (2015) also suggested a fully nonparametric method, where the propensity score is estimated nonparametrically using the NW-estimator with higher-order kernels. With careful selection of the kernel order and bandwidths, Abrevaya, Hsu, and Lieli (2015) argued that the fully nonparametric method could be more efficient than the semi-parametric approach. Nonetheless, this fully nonparametric method suffers from the curse of dimensionality, particularly when the sample size is insufficient, as the NW-estimator for the propensity score becomes less effective in high-dimensional settings, which are often required to satisfy the ignorability assumption.

Before Abrevaya, Hsu, and Lieli (2015), nonparametric estimation of the propensity score had already been applied to the estimation of the ATE, as reviewed by G. W. Imbens (2004). The potential of nonparametric propensity score estimators to improve efficiency and avoid model misspecification has been recognized, though their limitations in high-dimensional settings are also well known. To address these challenges, SDR (Xia et al., 2002) and the Nadaraya-Watson estimator (NW-estimator) have been employed in estimating treatment effect estimators such as ATE (Zhao et al., 2022; W. Luo and Y. Zhu, 2020; Huang and Chiang, 2017; Huang and Yang, 2020) and QTE (Y. Zhang et al., 2020).

For CATE, N. Zhou and Lixing Zhu (2021) studied the asymptotic behavior of IPW-based estimators from Abrevaya, Hsu, and Lieli (2015), incorporating SDR. L. Li, N. Zhou, and Lixing Zhu (2022) extended this approach by focusing on outcome regression method which is related but different from the IPW method. N. Zhou and Lixing Zhu (2021) using a similar set up with our study, but using the treatment indicator as the response variable for SDR, which may not be the most efficient choice in ATE estimation. The literature for ATE suggests that applying SDR based on the potential outcome would make the IPW estimator more efficient than using the treatment indicator (Hahn, 2004), making the choice of response in SDR a meaningful topic for CATE estiamtion.

Additionally, the exact impact of high dimensionality is not clearly articulated in the existing literature. Unlike the conventional NW-estimator, when a higher-order kernel is employed with its order increases with dimensionality, the effects of high dimensionality

become less discernible from the convergence rate from asymptotic properties. In our study, we will provide a detailed explanation of this phenomenon.

Although robust methods combined with SDR have been explored for estimating the ATE (Y. Zhang et al., 2020), there is currently no research addressing the construction of a fully nonparametric, robust CATE estimator that incorporates SDR. This gap in the literature highlights the need for further investigation.

Hence, in this study, we apply the concept of M-estimators to the fully nonparametric CATE estimator proposed by Abrevaya, Hsu, and Lieli (2015) and extend it with SDR to improve its robustness to high-dimensional covariates. The asymptotic theorems are derived, and a confidence interval is provided based on the asymptotic normality. We then verify the theoretical results with simulations. Based on the theoretical and simulation results, we observe that the CATE estimator combined with SDR behaves differently from the ATE estimator with SDR found in the existing literature. First, asymptotically, the convergence rate of the fully nonparametric CATE estimator is affected by the dimensionality of the covariates, whereas in the case of ATE, the convergence rate remains fixed, and only the asymptotic variance is typically considered. Second, the higher-order kernels used in our proposed estimator require a sufficiently large sample size to function effectively. These findings highlight the substantial influence of covariate dimensionality on estimator performance, underscoring the importance of dimension reduction.

Furthermore, we apply our proposed method to analyze data from the 2007-2008 NHANES to examine the impact of participation in the National School Lunch Program (NSLP) on Body Mass Index (BMI), focusing specifically on childrens age as a conditioning variable. Our analysis reveals a significant negative effect on BMI for children under 7 and a positive effect for those aged 8 and older. These findings, which contrast with the non-significant treatment effect reported by Huang and Chan (2017), highlight substantial heterogeneity across age groups and underscore the importance of using CATE estimators tailored to specific subpopulations for a comprehensive understanding of treatment effects in real-world applications.

The rest of the paper is organized as follows: Section 2 introduces the setup and motivation for our proposed estimator. In Section 3, we develop the theoretical properties of the

estimator without SDR, while Section 4 extends these results to include SDR and provides a confidence interval for inference purposes. Section 5 presents the simulation results, and Section 6 provides the data analysis.

4.2 Problem Setup and Motivation

4.2.1 Problem Setup and Notation

Following the potential outcomes framework Rubin (1974), we let T be an indicator of the treatment status with $T = 1$ for receiving treatment and $T = 0$ otherwise. Let X be a d-dimensional vector of covariates with $d \geq 2$ and Y be the response. Define $Y^{(1)}$ as the potential outcome corresponding to receiving treatment and $Y^{(0)}$ without treatment. As only one of the potential outcomes can be observed for any individual, the relationship between the observed outcome and potential outcomes can be written as $Y = TY^{(1)} + TY^{(0)}$.

The observed data denoted as $\{(T_{\mathrm{i}}, X_{\mathrm{i}}, Y_{\mathrm{i}})\}_{\mathrm{i}=1}^{n}$ is a random sample of size n from the joint distribution (T, X, Y).

Let $X_1 \in \mathbb{R}^l$ be a subvector of $X \in \mathbb{R}^d$ with $1 \leq l \leq d$. The CATE is defined as

$$\tau(x_1) = E[Y^{(1)} - Y^{(0)} \mid X_1 = x_1] = \mu_1(x_1) - \mu_0(x_1),$$

where $\mu_1(x_1) = E[Y^{(1)} \mid X_1 = x_1]$, $\mu_0(x_1) = E[Y^{(0)} \mid X_1 = x_1]$ are the conditional means of potential outcomes.

To make the unconfoundedness assumption hold, we need to include a large number of covariates to guarantee the necessary ones are included. Thus we may assume X to be a high dimensional vector and X_1 to be a relatively low dimensional subvector of X.

4.2.2 Identify Outlier-resistant CATE

In Chapter 2, we assumed the potential outcomes are heavy-tailed symmetric distributed around their conditional means. To provide an outlier-resistant estimator, we combined the inverse propensity weighting estimator (Abrevaya, Hsu, and Lieli, 2015) with the local

M-estimator (Härdle, 1984). For the convenience of applying the M-estimator, we estimate $\hat{\mu}_1(x_1)$ and $\hat{\mu}_0(x_1)$ separately and estimate the CATE by subtraction of the two estimators:

$$\hat{\tau}(x_1) = \hat{\mu}_1(x_1) - \hat{\mu}_0(x_1).$$

Similar to identifying the estimator of CATE, since only one of the potential outcomes can be observed on each of the observations, the unconfoundedness assumption (Rosenbaum and Rubin, 1983) and positivity assumption are frequently used in the estimation of treatment effect to identify the CATE estimator (Abrevaya, Hsu, and Lieli, 2015; W. Luo, Y. Zhu, and Ghosh, 2017).

Assumption (A1)

1. (Unconfoundedness) $(Y^{(1)}, Y^{(0)}) \perp T \mid X$.

2. (Positivity) Let $\boldsymbol{\pi}(x) = P(T = 1 \mid X = x)$ be the propensity score, and there exists $C > 0$ such that $P(C \leq \boldsymbol{\pi}(X) \leq 1 - C) = 1$.

Based on the definitions and (A1), $\mu_1(x_1)$ can be derived as the solution of estimating equation

$$0 = E\left[\frac{T}{\boldsymbol{\pi}(X)}(Y - \mu_1(x_1)) \;\mid\; X_1 = x_1\right],$$

And the estimator $\hat{\mu}_1(x_1)$ can be obtained by solving the empirical version of the estimating equation

$$0 = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1} - x_1}{h_1}\right)\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}(Y_{\mathrm{i}} - \hat{\mu}_1(x_1)),$$

where K_1 is a kernel function used in the NW-estimator, $\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})$ is the estimator of $\boldsymbol{\pi}(X_{\mathrm{i}}) = E(T \mid X_{\mathrm{i}})$. The other conditional mean $\hat{\mu}_0(x_1)$ can be estimated similarly.

To apply the robust method, as shown in our first study, we need the following assumption on the robust loss function $\psi(\cdot)$.

Assumption (A2) The robust function $\psi(\cdot)$ is antisymmetric and the conditional density $f(y^{(1)} \mid x_1)$ is symmetric with respect to $\mu_1(x_1)$.

With assumption (A2), we have

$$0 = E\left[\frac{T}{\boldsymbol{\pi}(X)}\psi(Y - \mu_1(x_1)) \mid X_1 = x_1\right]. \tag{4.1}$$

We will investigate the solution $\hat{\mu}_1(x_1)$ to the following empirical estimating equation

$$0 = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_\mathrm{i} - x_1}{h_1}\right)\frac{T_\mathrm{i}}{\hat{\boldsymbol{\pi}}(X_\mathrm{i})}\psi(Y_\mathrm{i} - \mu_1(x_1)),$$

with modifications to the kernel and the estimation of the propensity score. These modifications will be explained in the following sections.

4.2.3 Estimation of The Propensity Score

As discussed in Abrevaya, Hsu, and Lieli (2015), the propensity score $\hat{\boldsymbol{\pi}}(X_\mathrm{i})$ can be estimated using either parametric or nonparametric methods, leading to what are known as semiparametric and nonparametric estimators. Although the parametric approach carries the risk of model misspecification, it typically achieves a fast convergence rate, allowing the error from the propensity score estimation to diminish sufficiently, which can be asymptotically dominated by the limiting distribution of the CATE estimator.

On the other hand, while the nonparametric method avoids the pitfalls of model misspecification, it converges more slowly, making its estimation error more significant. Without modification, using a nonparametric estimator like the Nadaraya-Watson (NW) estimator can lead to the error in the propensity score estimation dominating the overall limiting distribution. However, by using a higher-order kernel and carefully selecting the kernel order and bandwidths, as suggested by Abrevaya, Hsu, and Lieli (2015), the nonparametric estimator can achieve better asymptotic efficiency compared to the semiparametric estimator, as noted in their Comment 3 following Theorem 2. This efficiency gain is evident in their simulation results for lower dimensions (e.g., $p = 2$), although the results in higher dimensions (e.g., $p = 4$) are mixed, reflecting the challenges posed by the curse of dimensionality.

Therefore, in the following sections, we will first extend our previous work in Chapter 2 by applying higher-order kernels within a fully nonparametric framework. We will then tackle the challenge of high dimensionality by integrating SDR techniques.

4.3 Outlier Resistant CATE with Higher-Order Kernel

4.3.1 Definition and Mechanism of Higher-Order Kernels

The higher-order kernel is defined as the order the first non-zero moment:

$\int u^k K(u)du$ equals 1 for$k = 1$, equals 0 for $k \in \{2, \cdots, s-1\}$, and in general not equals 0 when $k \geq s$.

Follows G. W. Imbens and Ridder (2009) and Sen (2011), we're able to construct the higher-order kernel based on an exisited symmetric kernel κ. In later sections, higher-order kernels with multiple dimensions will be used, especially in the estimation of propensity scores. As described in G. W. Imbens and Ridder (2009), multi-dimensional higher-order kernels can be constructed by taking products of higher-order univariate kernels.

The direct effect of a higher-order kernel on the NW-estimator is reducing the estimation bias caused by the curvature of the estimated function. To provide some intuitive understanding, the NW-estimator essentially computes a local average of the response values Y for observations near the conditioned value x. Consider a specific point x_0: if the true conditional mean function increases to the left of x_0 and decreases to the right, and if the observations are symmetrically distributed around the true mean, the local average of Y around x_0 will generally be less than the true conditional mean.

As mentioned by J. Marron (1994), a higher-order kernel has a taller peak at the center and assigns negative weights to the tails. This design achieves two important effects: it gives more weight to observations near the center (which reduces bias) and uses the negative weights on the tails to counteract the downward bias. Together, these adjustments bring the estimator closer to the true value.

When the bias is reduced, the bandwidth based on the optimal mean squared error (MSE) becomes the solution to $(nh)^{-1} = h^s$, leading to $h = n^{-\frac{1}{1+2s}}$. Therefore, although

the variance is not directly influenced by the kernel's order, a larger bandwidth is chosen to minimize the MSE, resulting in a higher convergence rate when using a higher-order kernel.

4.3.2 Asymptotic Properties

Intuitively, the application of a higher-order kernel in nonparametric estimation of the propensity score can achieve higher convergence rate, making its estimation error asymptotically negligible compared to the limiting distribution of $\hat{\mu}_1(x_1)$. However, reducing the error in propensity score estimation to such a degree requires using a kernel with very high order, which can introduce numerical and finite sample problems (as discussed in a later section). To address this issue, Abrevaya, Hsu, and Lieli (2015) shows that when higher-order kernels are applied to both the estimation of the propensity score and the estimation of CATE, and when certain relationships between the bandwidths and kernel orders are maintained, a more efficient estimator can be obtained with an acceptable level of kernel order, particularly when the dimension of covariates is small.

In this subsection, we derive the asymptotic properties of the outlier-resistant estimator when the propensity score is estimated using the leave-one-out version of the Nadaraya-Watson (NW) estimator:

$$\hat{\boldsymbol{\pi}}(X_{\mathrm{i}}) = \frac{\sum_{\mathrm{j}\neq\mathrm{i}} T_{\mathrm{j}} K\left(\frac{X_{\mathrm{j}}-X_{\mathrm{i}}}{h}\right)}{\sum_{\mathrm{j}\neq\mathrm{i}} K\left(\frac{X_{\mathrm{j}}-X_{\mathrm{i}}}{h}\right)}.$$

The estimator of $\mu_1(x_1)$ is then obtained by solving the estimating equation

$$0 = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}}-x_1}{h_1}\right)\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}\psi(Y_{\mathrm{i}}-\mu_1(x_1)).$$

Define

$$H_n(x_1,s) = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n}\psi(Y_{\mathrm{i}}-s)\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right),$$

and define

$$H(x_1,s) = E\left[\psi(Y_{\mathrm{i}}-s)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} \mid X_{\mathrm{i}1}=x_1\right] g(x_1).$$

With the conditions and notations listed in the appendix, we begin by demonstrating the consistency of the estimating equation.

Lemma 4.3.1. *Given assumption (A1) to (A8), we have* $H_n(x_1, s) \xrightarrow{P} H(x_1, s)$ *for each* $x_1 \in \chi$.

Based on the consistency of the estimating equation, the consistency of $\hat{\mu}_1(x_1)$ can be established similarly to our previous work or as discussed in Härdle (1984).

Theorem 4.3.1. *Suppose* $\psi(\cdot)$ *is continuous and strictly increasing, and there exists a constant* c *such that* $H(x_1, s) > 0$ *for* $s > c$ *and* $H(x_1, s) < 0$ *for* $s < c$. *Then, under the assumption (A1) to (A8),* $\hat{\mu}_1(x_1) \xrightarrow{P} \mu_1(x_1)$.

For continuous but non-monotone ψ functions, such as Tukey's biweighted loss, we define the estimator $\tilde{\mu}_1(x_1)$ as the value closest to $\hat{\mu}_1(x_1)$ among estimators derived from monotone ψ functions. Specifically, $\tilde{\mu}_1(x_1)$ is given by:

$$\mid \hat{\mu}_1(x_1) - \tilde{\mu}_1(x_1) \mid = \inf\{\mid s - \hat{\mu}_1(x_1) \mid : H_n(x_1, s) = 0\}.$$

The consistency of $\tilde{\mu}_1(x_1)$ can be demonstrated using the same approach as in our previous study Chapter 2.

With the consistency of $\hat{\mu}_1(x_1)$ established, we then proceed to prove its asymptotic normality.

Define

$$Z_n(x_1) = \frac{C_1(x_1)\left(\hat{\mu}_1(x_1) - \mu_1(x_1)\right)}{\left[\frac{\alpha_n(2)}{n}\sigma^2(x_1)g^{-1}(x_1)\right]^{1/2}},$$

where

$$\begin{aligned}
\sigma^2(x_1) = E\Bigg[\Bigg(\Big(\psi(Y_{\mathrm{i}} - \mu_1(x_1)) - E[\psi(Y_{\mathrm{i}}^{(1)} - \mu_1(x_1)) \mid X_{\mathrm{i}}]\Big)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} \\
+ E[\psi(Y_{\mathrm{i}}^{(1)} - \mu_1(x_1)) \mid X_{\mathrm{i}}]\Bigg)^2 \mid X_{\mathrm{i1}} = x_1\Bigg], \\
C_1(x_1) = E(\psi'(Y^{(1)} - \mu_1(x_1)) \mid X_1 = x_1).
\end{aligned}$$

Theorem 4.3.2. *Suppose the conditions in Theorem refThmP3:Consistency of mu1hat hold, we have* $Z_n \xrightarrow{D} N(0, 1)$ *if further satisfies the properties:*

(1) $\gamma_n = \int \mid \delta(u) \mid^{2+\eta} du < \infty$ *for some* $\eta > 0$,

(2) $\gamma_n = o(n^{\eta/2}\alpha_n(2)^{1+\eta/2})$ *as* $n \to \infty$.

Since by definition $\alpha_n(2) = O(h_1^l)$, the Theorem refThmP3:Normality of mu1hat indicates that $\hat{\mu}_1(x_1) - \mu_1(x_1)$ converges to normal distribution with convergence rate $\sqrt{nh_1^l}$ and asymptotic variance $\frac{\sigma^2(x_1)}{g(x_1)C_1(x_1)^2}$.

Remark 4.3.1. *In this remark, we verify that the property of asymptotic efficiency observed for the fully nonparametric estimator of CATE in Abrevaya, Hsu, and Lieli (2015) compared to the semiparametric estimator persists even when the estimator is generalized to a robust version.*

For simplicity, we denote $\psi^{(1)} = \psi(Y^{(1)} - \mu(x_1))$ *and* $\psi = \psi(Y - \mu(x_1))$. *Then, the asymptotic variance can be written as*

$$Var(\mu_1(x_1)) = \frac{\alpha_n(2)}{g(x_1)n} \cdot \frac{E\left[\left(\left(\psi - E[\psi^{(1)} \mid X]\right)\frac{T}{\pi(X)} + E[\psi^{(1)} \mid X]\right)^2 \mid X_1 = x_1\right]}{E\left[\frac{T}{\pi(X)}\psi'\left(Y - \mu_1(x_1)\right) \mid X_1 = x_1\right]^2}.$$

Focus on the numerator, we have

$$\begin{aligned} & E\left[\left(\left(\psi - E[\psi^{(1)} \mid X]\right)\frac{T}{\pi(X)} + E[\psi^{(1)} \mid X]\right)^2 \mid X_1 = x_1\right] \\ = \; & E\left[\psi^2 \frac{T}{\pi(X)^2} \mid X_1 = x_1\right] - E\left[E[\psi^{(1)} \mid X]^2\left(\frac{1}{\pi(X)} - 1\right) \mid X_1 = x_1\right] \\ \leq \; & E\left[\psi^2 \frac{T}{\pi(X)^2} \mid X_1 = x_1\right]. \end{aligned}$$

By Chapter 2, the term in the last row is the asymptotic variance when $\pi(X)$ *is estimated using a parametric method with a* $\sqrt{n}$ *convergence rate or when* $\pi(X)$ *is known.*

Therefore, the asymptotic variance in the nonparametric case is smaller than in the semiparametric case. This result is consistent with the findings for the non-robust CATE estimator in Abrevaya, Hsu, and Lieli (2015).

Remark 4.3.2. *In the original Nadaraya-Watson (NW) estimator with d dimensional covariates, the convergence rate of the bias is* $O(h^2)$ *and of the variance is* $O((nh^d)^{-1})$. *The*

mean squared error (MSE) optimal bandwidth is $h = O(n^{-\frac{1}{d+2}})$. *Increasing the dimension directly increases the variance. However, this effect is not evident in the asymptotic theorems when higher-order kernels are properly used. As illustrated in Comment 6 of Abrevaya, Hsu, and Lieli (2015), the relationship between the dimension and the order of the kernels is suggested to be* $s = d$ *for* d *even,* $s = d+1$ *for* d *odd, and* $s_1 = s+2$ *to ensure the conditions are met.*

For our estimator $\hat{\mu}_1(x_1)$, *the bias is* $O(h^{d+2})$ *and the variance is* $O((nh)^{-1})$. *With the MSE optimal bandwidth, the variance is* $O(n^{\frac{d}{d+2}})$, *which increases with the dimension. Even though the constant part of* $\alpha_n(2)$ *may increase with the order of the kernel, an estimator with higher dimensions will have a smaller overall variance compared to one with lower dimensions when the sample size is sufficiently large. Nonetheless, the curse of dimensionality is observed in simulation results. This indicates that, unlike the original NW estimator, the curse of high-dimensionality does not fully manifest in the asymptotic distribution of our proposed estimator.*

This observation motivates a review of the literature to explore the potential drawbacks of higher-order kernels.

First of all, as pointed out by Hardle (1986), higher-order kernels pay the price in terms of increased variance. This phenomenon is indeed reflected in our asymptotic properties in the norm of kernel but does not fully account for the increase in the convergence rate.

Although we haven't found a theoretical explanation, some observations in J Steve Marron and Wand (1992) might help explain why higher-order kernels often perform poorly in finite samples. Marron's simulations suggest that whether it's proper to use a higher-order kernel depends on the sample size. He showed that while there is usually a sample size large enough for higher-order kernels to be clearly better, this required size is often very large for density estimation. Additionally, Marron found that as the kernel order increases, the benefits become smaller, regardless of the sample size.

Additionally, an intuitive explanation is provided in J. Marron (1994). Marron suggested that higher-order kernels are more effective than nonnegative ones when the underlying function being estimated closely resembles a parabola within neighborhoods of a radius roughly equivalent to the effective window width of the higher-order kernel. However, for small sam-

ple sizes, the mean squared error (MSE) optimal bandwidth is usually too large, which can result in the underlying function within the bandwidth-controlled neighborhood not approximating a parabola. For instance, multiple zigzags may occur within the neighborhood.

4.3.3 A Simple Example of the Effects of Dimensionality on Higher-Order Kernels

In this subsection, we use a simple example to verify our findings from the previous subsection.

Specifically, we apply the Nadaraya-Watson (NW) estimator with higher-order kernels to estimate the model:

$$y = 10 + 5x + \cos(10x) + \epsilon,$$

where $x \sim \text{Unif}(-0.5, 0.5)$ and $\epsilon \sim N(0, 0.25)$.

We estimate $E[Y \mid X = x]$ using higher-order normal kernels with orders $s \in \{2, 4, 6, 8, 10\}$, while varying the sample size n from 100 to 5000.

For each combination of kernel order and sample size, we estimate and output the Mean Integrated Squared Error (MISE) by

$$\widehat{\text{MISE}} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{k}\sum_{\text{i}=1}^{k}(\hat{f}_r(x_k) - f(x_k))^2,$$

with $R = 500$ repetitions.

Here, x_k are equally spaced design points between -0.5 and 0.5, and $\hat{f}_r(x_k)$ denotes the estimator of $f(x_k)$ using the data generated in the r-th repetition.

From the simulation results, we observe that there is indeed a requirement for the sample size when using higher-order kernels. As shown in the graph, higher-order kernels need a larger sample size to outperform those with lower orders. Furthermore, the required sample size increases dramatically with the kernel order. This observation is consistent with the findings of J Steve Marron and Wand (1992).

Since, the relationship between the dimension and the order of the kernels is suggested to be $s = d$ for d even, $s = d + 1$ for d odd, and $s_1 = s + 2$, for our proposed estimator,

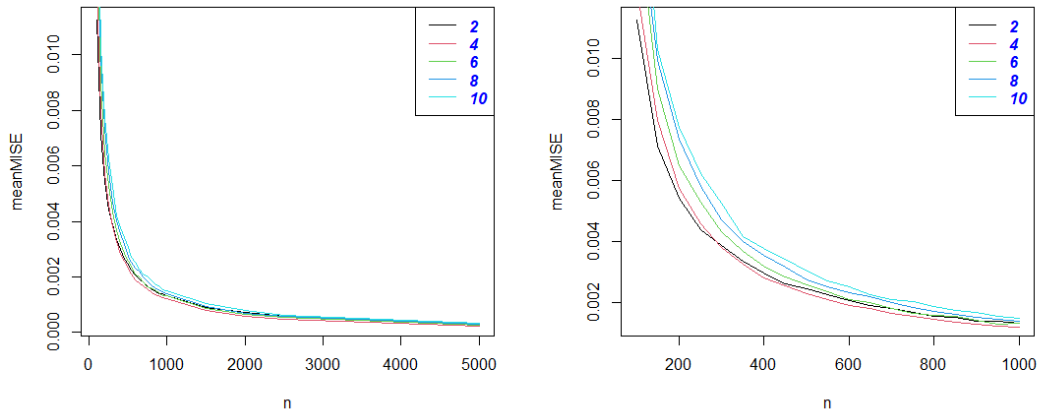

Figure 4.1. Average MISE of kernel regression with different kernel orders across various sample sizes.

when the dimension of covariates exceeds 10, an order of 12 may be necessary for estimating $\mu_1(x_1)$. This could require 4000-5000 observations to ensure that the higher-order kernel outperforms one without higher-order kernels. Thus, it is reasonable to consider dimension reduction techniques for the fully nonparametric estimator of CATE when the sample size is limited.

4.4 Nonparametric Outlier-resistant CATE estimator with SDR

4.4.1 Central Mean space

In the previous sections, we emphasized the necessity of dimension reduction without specifying the criteria to use. In this section, we explore adopting the concept of SDR, where the covariates X are reduced through a linear transformation and replaced by $B^T X$, with B being the transformation matrix. Then the estimating equation 4.1 can be rewritten as:

$$0 = E\left[\frac{T}{\pi(B^T X)}\psi(Y - \mu_1(x_1)) \mid X_1 = x_1\right]. \tag{4.2}$$

However, the solution defined by (4.2) is different from (4.1) in general.

The choices of the Central Subspace defined by B for the ATE estimator have been well studied, which may provide clues on possible choices for our case. Since the covariates are directly used in the estimation of the propensity score, it seems quite natural to reduce the dimension of X by finding a matrix B such that $S_T = B_T^T X$ have the smallest dimension satiesfy $T \perp X \mid B^T X$. Unfortunately, Hahn (1998) indicated that in the estimation of ATE, using such an S_T provides no improvement in the asymptotic variance over using the entire X. Instead, Hahn (2004) suggested finding a dimension reduced vector $S_{Y^{(0)},Y^{(1)}} = B^T_{Y^{(0)},Y^{(1)}} X$ with the smallest dimension such that $(Y^{(0)}, Y^{(1)}) \perp X \mid S_{Y^{(0)},Y^{(1)}}$ and showed it more asymptotically efficient than the estimator with X unless $S_{Y^{(0)},Y^{(1)}} = X$. De Luna, Waernbaum, and Richardson (2011) further considered an S_{min} which removes the components in $S_{Y^{(0)},Y^{(1)}}$ that are unrelated to T, that is $T \perp S_{Y^{(0)},Y^{(1)}} \mid S_{min}$. However, it is proved that the estimator using S_{min} is no more efficient than the estimator with $S_{Y^{(0)},Y^{(1)}}$ (Brookhart et al., 2006; Shortreed and Ertefaie, 2017). Y. Zhang et al. (2020) studied SDR under the quantile treatment effect (QTE). Similar to our proposed method, the estimator of QTE is estimated by the difference of two separate quantile estimators of potential outcomes. The SDR subspaces S_{Y_k} for $k \in \{0, 1\}$ are estimated to satisfy $Y_k \perp X \mid S_{Y_k}$ and have the smallest dimensions.

In addition, since the CATE is defined as a conditional mean function, we can formulate a subspace by considering the independence of variables in terms of the conditional means (Cook and B. Li, 2002; Ma and Liping Zhu, 2012; W. Luo, Y. Zhu, and Ghosh, 2017), instead of focusing on the entire distributions. For example, instead of applying SDR based on the conditional independence $Y \perp X \mid B^T X$, we consider a weaker criterion: $E(Y \mid X) = E(Y \mid B^T X)$, which defines the central mean space of equal or smaller dimension.

With this background, we start by finding the central mean spaces to identify the CATE estimator based on dimension-reduced covariates. We will then compare these estimators to achieve higher efficiency for the new estimator.

4.4.2 Identify $\mu_1(x_1)$ in Central Mean Space

To find the central mean space that enables us to identify our estimator μ_1, suppose the dimension-reduced variables are (B, T, Y). To identify the estimating equation for the IPW method, we need the unconfoundedness and positivity assumptions. These assumptions can be rewritten in their dimension-reduced version as follows:

Assumption (A9):

1)$E[Y^{(1)}T \mid B^TX] = E[Y^{(1)} \mid B^TX]E[T \mid B^TX]$.

2) $E[T \mid B^TX]$ is bounded away from 0 and 1.

Except for (A9), the iterative law of expectation is also a key step in rewriting the potential outcome by the observed one. Unlike the original case, the set of covariates B^TX no longer generally includes X_1, and hence we need to assume:

Assumption (A10) $E[Y^{(1)} \mid X_1 = x_1] = E[E[Y^{(1)} \mid B^TX]X_1 = x_1]$ so that we can apply a similar approach to the iterative law of expectation in identifying our proposed estimator.

Theorem 4.4.1. *With assumption (A9) and (A10), we are able to identify $\mu_1(x_1)$ as the solution of*

$$0 = E\left[(Y^{(1)} - \mu_1(x_1))\frac{T}{E[T \mid B^TX]} \mid X_1 = x_1\right].$$

The identification of $\mu(x_1)$ is similar to the case without SDR:

To derive the estimator $\mu_1(x_1)$ based on the dimension-reduced variables, we start with:

$$\begin{aligned}
\mu_1(x_1) &= E[Y^{(1)} \mid X_1 = x_1] \\
0 &= E[Y^{(1)} - \mu_1(x_1) \mid X_1 = x_1] \\
0 &= E\left[E[Y^{(1)} - \mu_1(x_1) \mid B^TX] \mid X_1 = x_1\right] \\
0 &= E\left[E[Y^{(1)} - \mu_1(x_1) \mid B^TX]\frac{E[T \mid B^TX]}{E[T \mid B^TX]} \mid X_1 = x_1\right] \\
0 &= E\left[E[(Y^{(1)} - \mu_1(x_1))\frac{T}{E[T \mid B^TX]} \mid B^TX] \mid X_1 = x_1\right] \\
0 &= E\left[(Y^{(1)} - \mu_1(x_1))\frac{T}{E[T \mid B^TX]} \mid X_1 = x_1\right].
\end{aligned}$$

When B is defined by the lowest dimension that satisfies $E[Y^{(1)} \mid X] = E[Y^{(1)} \mid B^T X]$, assumption (A10) holds directly. However, in other cases, such as $E[T \mid X] = E[T \mid B^T X]$, (A10) may not hold in general. In these situations, we perform dimension reduction on covariates excluding X_1, ensuring that X_1 will be automatically included as a subvector of $B^T X$.

For assumption (A9), regarding the positivity assumption, it is commonly assumed to hold directly in theoretical work, while in real data applications, we may trim the estimated propensity score. For the unconfoundedness assumption, existing literature on SDR under ATE or QTE demonstrates that their criteria hold with dimension-reduced covariates. We follow this approach and provide a proof for a central mean space version instead of the whole distribution version found in the literature.

The problem then becomes: Given the unconfoundedness assumption in (A1), can we derive the dimension-reduced version of unconfoundedness in (A2) using the criteria for SDR?

4.4.3 Central Mean Space

In this subsection, we examine criteria from the existing literature on applying SDR to ATE and QTE estimations to determine if they preserve the unconfoundedness assumption after dimension reduction. Specifically, we consider two types of dimension reduction criteria:

1. $E[Y^{(1)} \mid X] = E[Y^{(1)} \mid S_{Y_1} = B^T X]$,
2. $E[T \mid X] = E[T \mid S_T = B^T X]$.

As we demonstrate in the next theorem, the dimension-reduced version of unconfoundedness can be established using either of the two criteria.

Theorem 4.4.2. *Suppose we have* $E[Y^{(1)} \mid X] = E[Y^{(1)} \mid B^T X]$ *or* $E[T \mid X] = E[T \mid B^T X]$. *Then, together with assumption (A1), we have*

$$E(Y^{(1)} T \mid B^T X) = E(Y^{(1)} \mid B^T X) E(T \mid B^T X).$$

4.4.4 Estimating The Central Mean Space by Modified rMAVE under Counterfactual Framework

For estimating the central mean subspace, several methods are available in the literature, such as inverse regression techniques proposed by Cook and B. Li (2002) and the MAVE method developed by Xia et al. (2002) and Xia (2007). In this study, we first employ MAVE and explore the modifications for applying it within a counterfactual framework. then we consider applying rMAVE íek and Härdle (2006), W. Yao and Q. Wang (2013), and Jing Zhang, Q. Wang, et al. (2021) to further incerase its robustness against heavy-tailed responses.

For S_T in this study, we can apply the MAVE method with T as the response and X as the covariates. However, for $Y^{(1)}$ and $Y^{(0)}$, we cannot directly obtain $S_{Y^{(1)}}$ and $S_{Y^{(0)}}$ directly because we do not observe $Y^{(1)}$ and $Y^{(0)}$ for each observation.

Let $\Omega(W)$ denote the support of W and $\Omega(W \mid V)$ denote the support of W given V. According to Theorem 1 in W. Luo, Y. Zhu, and Ghosh (2017), if

$$\Omega(B_t^T X) = \Omega(B_t^T X \mid T = t), \quad (t = 0, 1),$$

Then the central mean space estimated condition on $T = 1$ or $T = 0$ is the same as the unconditioned central mean space, and hence we can apply method like MAVE conditional on $T = 1$ or $T = 0$ when the response is set to $Y^{(1)}$ or $Y^{(0)}$.

The MAVE method is based on minimizing

$$E[\sigma_B^2(B^T X)] = E[E[[Y^{(1)} - E(Y^{(1)} \mid B^T X)]^2 \mid B^T X]].$$

In W. Luo, Y. Zhu, and Ghosh (2017), they adjust the MAVE algorithm for their estimator for the ATE and estimate the $\sigma_B^2(B^T X)$ by:

$$\hat{\sigma}_B^2(B^T X) = min_{a,b} \Sigma_{\mathrm{i}=1}^n [y_{\mathrm{i}} - a - b^T B^T (X_{\mathrm{i}} - X)]^2 W_{\mathrm{i}0},$$

where $W_{\mathrm{i}0} = \frac{T_{\mathrm{i}} K_h\{B^T(X_{\mathrm{i}} - Xx)\}}{\Sigma_{\mathrm{i}=1}^n T_{\mathrm{i}} K_h\{B^T(X_l - X)\}}$.

And $\hat{B}$ can be estimated by

$$\underset{B:B^TB=I}{\operatorname{argmin}} \left\{ \sum_{\mathrm{j}=1}^{n} \hat{\sigma}_B^2 \left(B^T X_{\mathrm{j}} \right) \right\}.$$

As we mentioned in Assumption (A10), applying SDR with T as the response will likely remove the interaction between the potential outcomes $Y^{(1)}$ and X_1. Therefore, as illustrated in the previous subsection, it is necessary to retain X_1 and apply the dimension reduction method to the remaining covariates. To achieve this, we define

$$B = \begin{bmatrix} 1 & 0^T \\ 0 & B_{-1} \end{bmatrix},$$

where 1 denotes an $l \times l$ identity matrix, 0 denotes an $l \times 1$ column vector of zeros, and B_{-1} is a $d \times d$ matrix. In the step of estimating B, we minimize the loss function by adjusting the values of B_{-1} while keeping the other elements fixed. This modification ensures that Assumption (A10) holds and guarantees that the estimator is correctly identified.

Since $Y^{(1)}$ is assumed to be heavy-tailed distributed, we may further adjust the loss function for rMAVE similar to íek and Härdle (2006), W. Yao and Q. Wang (2013), and Jing Zhang, Q. Wang, et al. (2021):

$$\hat{\sigma}_B^2(B^T X) = \min_{a,b} \sum_{\mathrm{i}=1}^{n} \rho \left[y_{\mathrm{i}} - \{a + b^T B^T (X_{\mathrm{i}} - X)\} \right] W_{\mathrm{i}0},$$

where

$$W_{\mathrm{i}0} = \frac{T_{\mathrm{i}} K_h \{B^T (X_{\mathrm{i}} - X)\}}{\sum_{\mathrm{i}=1}^{n} T_{\mathrm{i}} K_h \{B^T (X_{\mathrm{i}} - X)\}}.$$

4.4.5 Fully Nonparametric Outlier Resistant CATE with SDR

The results in Huang and Chan (2017) and W. Luo, Y. Zhu, and Ghosh (2017) indicate that, under reasonable conditions, the MAVE and the rMAVE estimator converges at a sufficiently fast rate. For simplicity, we will adopt the following assumption based on their findings we assume:

Assumption (A11) $\|\hat{B} - B\| = o_p(n^{-1/4})$.

This assumption ensures that the remainder term of the Taylor expansions in the proof of the following theorem are bounded.

Theorem 4.4.3. *Suppose that the conditions in Theorem 4.3.2 and assumptions (A9) to (A11) hold. Then, the estimator $\hat{\mu}_1(x_1)$ is asymptotically normal with mean $\mu_1(x_1)$ and variance given by*

$$Var(\hat{\mu}_1(x_1)) = \frac{\alpha_n(2)\sigma_B^2(x_1)}{ng(x_1)C_B(x_1)^2},$$

where

$$\sigma_B^2(x_1) = E\left[\left((\psi(Y - \mu_1(x_1)) - E[\psi(Y^{(1)} - \mu_1(x_1)) \mid B^T X])\frac{T}{\boldsymbol{\pi}(B^T X)} + E[\psi(Y^{(1)} - \mu_1(x_1)) \mid B^T X]\right)^2 \mid X_1 = x_1\right],$$

and

$$C_B(x_1) = E\left[\frac{T}{\boldsymbol{\pi}(B^T X)}\psi'(Y - \mu_1(x_1)) \mid X_1 = x_1\right].$$

4.4.6 Comparison of The Asymptotic Variances

In this subsection, we discuss the efficiency of our estimator with covariates X and dimension-reduced covariates $S = B^T X$. When different responses are used in the rMAVE, the matrix B will be estimated differently, leading to different asymptotic variances for the estimator of $\mu_1(x_1)$. For comparison, we denote the dimension-reduced covariates corresponding to responses Y and T as S_Y and S_T, respectively.

With different responses used in SDR, the convergence rate $\sqrt{nh_1}$ can achieve the optimal rate when h_1 minimizes the mean squared error (MSE) under a sufficiently large kernel order s_1, and it eventually approaches $\sqrt{n}$ as s_1 increases to infinity. Therefore, in the asymptotic

case as $n \to \infty$, the variance of the limiting distribution becomes the dominant factor influencing the estimator's performance.

Based on the asymptotic property in the last subsection, we have the asymptotic variance as

$$Var(\hat{\mu}_1(x_1)) = \frac{\alpha_n(2)}{ng(x_1)} \frac{\sigma^2(x_1)}{C_1^2(x_1)}.$$

The nonconstant part in the expression of the variance can be rewritten as

$$V_X(x_1) = \frac{E\left[\frac{Var(\psi(Y-\mu_1(x_1))|B^TX)}{\pi(B^TX)} \mid X_1 = x_1\right] + Var\left(E[\psi(Y^{(1)} - \mu_1(x_1)) \mid B^TX] \mid X_1 = x_1\right)}{E[\psi'(Y^{(1)} - \mu_1(x_1)) \mid X_1 = x_1]^2}.$$

Specifically, we have $B = I$ and $B^TX = X$ when SDR is not applied. By rewriting the observed response Y as $Y^{(1)}$, the denominator becomes independent of X. Therefore, for comparison, we only need to focus on the numerator $\sigma^2(x_1)$.

Denote $a = \psi(Y^{(1)} - \mu_1(x_1))$, we have

$$\begin{aligned}
& \sigma_X^2(x_1) - \sigma_S^2(x_1) \\
= \; & E\left[\frac{\mathrm{Var}(\psi(Y^{(1)} - \mu_1(x_1)) \mid X)}{\boldsymbol{\pi}(X)} \mid X_1 = x_1\right] + \mathrm{Var}\left(E[\psi(Y^{(1)} - \mu_1(x_1)) \mid X] \mid X_1 = x_1\right) \\
& -E\left[\frac{\mathrm{Var}(\psi(Y^{(1)} - \mu_1(x_1)) \mid X)}{\boldsymbol{\pi}(S)} \mid X_1 = x_1\right] - \mathrm{Var}\left(E[\psi(Y^{(1)} - \mu_1(x_1)) \mid S] \mid X_1 = x_1\right) \\
= \; & E\left[\left(\frac{1}{\boldsymbol{\pi}(X)} - 1\right) \mathrm{Var}(a \mid X) \mid X_1 = x_1\right] - E\left[\left(\frac{1}{\boldsymbol{\pi}(S_T)} - 1\right) \mathrm{Var}(a \mid S) \mid X_1 = x_1\right].
\end{aligned}$$

For S_T, we have $\boldsymbol{\pi}(X) = \boldsymbol{\pi}(S_T)$, then by law of total variance we have

$$\begin{aligned}
\sigma_X^2(x_1) - \sigma_S^2(x_1) &= E[Var(\sqrt{\boldsymbol{\pi}(X)^{-1} - 1}a \mid X)] - E[Var(\sqrt{\boldsymbol{\pi}(X)^{-1} - 1}a \mid S_T)] \\
&= -Var(E(\sqrt{\boldsymbol{\pi}(X)^{-1} - 1}a \mid X)) + Var(E(\sqrt{\boldsymbol{\pi}(X)^{-1} - 1}a \mid S_T)) \\
&\quad + Var(\sqrt{\boldsymbol{\pi}(X)^{-1} - 1}a) - Var(\sqrt{\boldsymbol{\pi}(X)^{-1} - 1}a) \\
&= -E[Var(E[\sqrt{\boldsymbol{\pi}(X)^{-1} - 1}a \mid X] \mid S_T)].
\end{aligned}$$

Hence we have $\sigma_X^2(x_1) \leq \sigma_T^2(x_1)$.

Then we compare $\sigma_X^2(x_1)$ and $\sigma_Y^2(x_1)$. With S_Y, we have $E[a \mid S_Y] = E[a \mid X]$, then we have

$$\sigma_X^2(x_1) - \sigma_Y^2(x_1) = E\left[\frac{\text{Var}(a \mid X)}{\boldsymbol{\pi}(X)} \mid X_1 = x_1\right] - E\left[\frac{\text{Var}(a \mid S_Y)}{\boldsymbol{\pi}(S_Y)} \mid X_1 = x_1\right].$$

By Jensen's inequality,

$$\begin{aligned}
E\left[\frac{\text{Var}(a \mid X)}{\boldsymbol{\pi}(X)}\right] &= E\left[E\left[\frac{\text{Var}(a \mid X)}{\boldsymbol{\pi}(X)} \mid S_Y\right]\right] \\
&= E\left[E\left[\frac{\text{Var}(a \mid S_Y)}{\boldsymbol{\pi}(X)} \mid S_Y\right]\right] \\
&= E\left[\text{Var}(a \mid S_Y) E\left[\frac{1}{\boldsymbol{\pi}(X)} \mid S_Y\right]\right] \\
&\geq E\left[\text{Var}(a \mid S_Y) \frac{1}{E[\boldsymbol{\pi}(X) \mid S_Y]}\right] \\
&= E\left[\frac{\text{Var}(a \mid S_Y)}{\boldsymbol{\pi}(S_Y)}\right].
\end{aligned}$$

Thus, we have $\sigma_X^2(x_1) \geq \sigma_Y^2(x_1)$.

Combining the two comparisons, we establish the relationship between the asymptotic variances: $\sigma_T^2(x_1) \geq \sigma_X^2(x_1) \geq \sigma_Y^2(x_1)$. This implies that with higher-order kernels, when the same order of kernel is used across all cases (determined by the sample size), it is preferable to use Y as the response for SDR rather than T due to the smaller asymptotic variance. This finding is consistent with the results presented in Y. Zhang et al. (2020) and De Luna, Waernbaum, and Richardson (2011).

4.4.7 The Asymptotic Distribution of $\hat{\tau}(x_1)$

Since $\hat{\mu}_1(x_1)$ and $\hat{\mu}_0(x_1)$ are consistent estimators of $\mu_1(x_1)$ and $\mu_0(x_1)$, respectively, we have under conditions of Theorem 4.4.3:

$$\hat{\tau}(x_1) = \hat{\mu}_1(x_1) - \hat{\mu}_0(x_1) \xrightarrow{P} \mu_1(x_1) - \mu_0(x_1) = \tau(x_1).$$

Then similar to the proof of normality of $\hat{\mu}_1(x_1)$, we have the normality:

$$\sqrt{\frac{ng(x_1)}{\alpha_n(2)\sigma^2}}(\hat{\tau}(x_1)-\tau(x_1)) \xrightarrow{D} N(0,1),$$

where

$$\begin{aligned}
\sigma^2 &= E\left[\left(\frac{H^{*1}(x_1)}{D^1(x_1)}-\frac{H^{*0}(x_1)}{D^0(x_1)}\right)^2 \mid X_1=x_1\right] \\
&= \frac{\alpha_n(2)}{ng(x_1)}\left(E\left[\frac{\mathrm{Var}(\psi(Y^{(1)}-\mu_1(x_1)))}{\boldsymbol{\pi}(X)D^1(x_1)} \mid X_1=x_1\right]\right. \\
&\quad \left.+E\left[\frac{\mathrm{Var}(\psi(Y^{(0)}-\mu_0(x_1)))}{(1-\boldsymbol{\pi}(X))D^0(x_1)} \mid X_1=x_1\right]\right) \\
&\quad +\mathrm{Var}\left(\frac{E[\psi(Y^{(1)}-\mu_1(x_1)) \mid X]}{D^1(x_1)}-\frac{E[\psi(Y^{(0)}-\mu_0(x_1)) \mid X]}{D^0(x_1)} \mid X_1=x_1\right),
\end{aligned}$$

$$H^{*1}=\left(\psi(Y_{\mathrm{i}}-\mu_1(x_1))-E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1)) \mid X_{\mathrm{i}}]\right)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}+E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1)) \mid X_{\mathrm{i}}], \quad (4.3)$$

$$H^{*2}=\left(\psi(Y_{\mathrm{i}}-\mu_0(x_1))-E[\psi(Y_{\mathrm{i}}^{(0)}-\mu_0(x_1)) \mid X_{\mathrm{i}}]\right)\frac{1-T_{\mathrm{i}}}{1-\boldsymbol{\pi}(X_{\mathrm{i}})}+E[\psi(Y_{\mathrm{i}}^{(0)}-\mu_0(x_1)) \mid X_{\mathrm{i}}], \quad (4.4)$$

$$D^1(x_1)g(x_1)=E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'(Y_{\mathrm{i}}-\mu_1(x_1)) \mid X_{\mathrm{i}1}=x_1\right]g(x_1),$$

$$D^0(x_1)g(x_1)=E\left[\frac{1-T_{\mathrm{i}}}{1-\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'(Y_{\mathrm{i}}-\mu_0(x_1)) \mid X_{\mathrm{i}1}=x_1\right]g(x_1).$$

The proof of the above result can be found in the appendix.

4.5 Simulation

4.5.1 Data Generation Process

In our paper, we consider 3 models:

model 1: $Y^{(1)}=(X_1+0.5\mathrm{e}xp(-X_1^2/0.1)X_2X_3X_4+v$,

and $\boldsymbol{\pi}(X)=\Lambda(0.5(X_1+0.5\mathrm{e}xp(-X_1^2/0.1)+2(X_2+X_3+X_4))$.

In this model, we expect S_t to be more effective in reducing the curse of dimensionality with dimension 1 while S_Y is less effective with dimension 4.

model 2: $Y^{(1)} = (X_1 + 0.5exp(-X_1^2/0.1) + 2(X_2 + X_3 + X_4) + v$ and $\boldsymbol{\pi}(X) = \Lambda((X_1 + 0.5exp(-X_1^2/0.1))X_2X_3X_4)$.

In the second model, S_Y have dimension 1 while S_t has dimension 4.

model 3: $Y^{(1)} = (X_1 + 0.5exp(-X_1^2/0.1) + 2(X_2 + X_3 + X_4) + v$ and $\boldsymbol{\pi}(X) = \Lambda(0.5(X_1 + 0.5exp(-X_1^2/0.1) + 2(X_2 + X_3 + X_4))$

In the third model, we expect S_Y and S_t to be both useful in dimension reduction with dimension 1.

In those models, we consider generate $X = (X_1, X_2, X_3, X_4)$ by $X_1 = \mathrm{e}_1, X_2 = (1+2X_1)+\mathrm{e}_2, X_3 = (1 + 2X_1) + \mathrm{e}_3, X_4 = (-1 + X_1)^2 + \mathrm{e}_4$ with $\mathrm{e_j} \sim \mathrm{unif}(-0.5, 0.5)$ for $\mathrm{j} = 1, 2, 3, 4$, which is the same as the 4 covariates model in Abrevaya, Hsu, and Lieli (2015). Except for that, we also consider models with zero coefficients follow the high-dimensional models in N. Zhou and Lixing Zhu (2021): $X_\mathrm{j} =\mid 1 + 1/(11 - \mathrm{j})X_1 \mid - \mid 1 + 1/\mathrm{je_j} \mid$ for $3 < \mathrm{j} \leq 9$; $X_\mathrm{j} =\mid 1 + 1/(11)X_1 \mid - \mid 1 + 1/10\mathrm{e_j} \mid$ for $\mathrm{j} = 10$; also with $\mathrm{e_j} \sim \mathrm{unif}(-0.5, 0.5)$ for $\mathrm{j} = 5, 6, ..., 10$.

For the error term v, we consider: $v = \epsilon v_1 + (1 - \epsilon)v_2$ where $v_1 \sim N(0, 0.25^2)$, and $v_1 \sim 2 \cdot T3$. The contamination ratio ϵ is a binary variable with values $\{0, 0.2, 0.4\}$.

4.5.2 Simulations for Outlier Resistant Estimator

Based on the data-generating process mentioned in the previous subsection, we investigate the finite sample performance of our proposed estimators. With the derivative of loss function $\psi(x) = x$, we have:

IPWN: the propensity score is estimated by NW-estimator without dimension reduction.

IPWP: the propensity score is estimated by logistic regression.

IPWSy: the propensity score is estimated by the MAVE method with Y as the response.

IPWSt: the propensity score is estimated by the MAVE method with T as the response.

Similarly, in the case $\psi(x)$ is the derivative of the Huber loss, we have rIPWN, rIPWP, rIPWSy, rIPWSt denote estiamtors based on the Huber loss with the corresponding types of propensity score estimator. In addition, we ues rIPWSry to denote the case when robust rMAVE is used for the dimension reduction.

We apply the methods on data set generated in the last subsection, with sample size 200 and 500 repititions on 37 equal distance points between $x_1 = -0.45$ to $x_1 = 0.45$. The MISE, standard error, mean of the width of confidence intervals and the mean of coverage rate of confidence intervals are printed.

In the rMAVE method, we use the normal kernel for estimation. The bandwidths and dimension are chosen by cross-validation. For the nonparametric estimation of the propensity scores, we employ the Epanechnikov kernel to satisfy the necessary assumptions. However, in the estimation of the $\hat{\mu}_1(x_1)$, we opt for the normal kernel because it is less sensitive to bandwidth selection and produces continuous estimators.

Model evaluation and selection also play a key role in our proposed estimator. However, under the counterfactual framework, the potential outcomes are not observed directly on each of the observations, and we are not able to calculate commonly used loss metrics such as mean squared error(MSE) directly.

In rMAVE, as we discussed, based on the theorem 1 in W. Luo, Y. Zhu, and Ghosh (2017), we can do cross-validation based on a grid search through the estimators of $E[(Y-\mu_1(x_1))^2 \mid T = 1]$ and $E[(Y - \mu_0(x_1))^2 \mid T = 0]$. For the bandwidth selection after the dimension reduction, the bandwidths are selected as $h = a \cdot n^{\frac{-1}{k+s+1}}$ and $h_1 = a_1 \cdot n^{\frac{-1}{l+2s_1-1}}$ with the constant a selected by regular cross-validation for the selection of bandwidth for the estimation of the propensity score and a_1 selected by a causal cross-validation (P. Gutierrez and Gérardy, 2017) for the estimation of $\mu_1(x_1)$.

4.5.3 Computation Problems

In my simulations, I faced three computation problems:

1. When the dimension of the covariates is high(e.g. $p > 12$), we need to use kernels with very large orders to estimate the propensity score and the CATE. Follow the steps in G. W. Imbens and Ridder (2009), when the order is large, the matrix need to be inverted will have a large size, and some of the terms in the last row/column will be extremely large and hence the matrix will be computationally unstable due to singular or multicollinearity.

2. Abrevaya, Hsu, and Lieli (2015) suggests the construction of multi-dimensional kernels by multiplying higher order kernels for each dimension together, this can also lead the value of the kernels to exceed the lower limit of values allowed in R. Although the problem can be alleviated by some simplification and scaling, we are only to make it possible to construct kernels with dimensions over 10 but less than 15 due to the singularity and multicollinearity problems.

3. When the dimension of the covariates is large and the bandwidth is smaller than necessary, there might not be enough data points in some specific window and errors will be created. This creates some difficulty when doing cross-validation. Since this is exactly due to the curse of high dimensionality, we chose greater bandwidths directly when the MSE is not able to calculate for this reason.

4.5.4 Simulation Result

We highlight several aspects of the simulation results reported in Tables 1 through 3.

1. Comparing between robust and non-robust methods.

The robust methods perform similarly to their non-robust counterparts in cases where there are no outliers. However, in contaminated cases, the robust methods effectively reduce the MISE, and the widths of the confidence intervals are also significantly reduced.

2. Model misspecification.

Due to model misspecification, the performance of IPWP is unstable. Unlike the results in Abrevaya, Hsu, and Lieli (2015), our simulation results show that IPWP does not outperform IPWN in many cases, and sometimes performs worse than other methods, especially in Models 1 and 3. However, IPWN exhibits more consistent performance across all three models.

3. Comparing the effect of dimension reduction in the non-robust case.

For $d = 4$, IPWSt and rIPWSt do not achieve lower MISE than IPWN in all three models. In our theoretical analysis, IPWSt is shown to have greater asymptotic variance than IPWN, which may lead to higher overall MISE, despite alleviating the curse of dimensionality after dimension reduction. On the other hand, IPWSy has lower MISE than IPWN only in Models

Table 4.1. The distribution of $\hat{\tau}(x_1)$ for Model 1.

	d=4						d=10					
Method	MISE	Bias	SD	SE	Width	Rate	MISE	Bias	SD	SE	Width	Rate
OutlierRatio=0												
IPWN	0.0025	0.0018	0.0480	0.0550	0.2158	0.9633	0.0035	0.0029	0.0455	0.0630	0.2469	0.9738
rIPWN	0.0025	0.0013	0.0480	0.0548	0.2150	0.9627	0.0035	0.0029	0.0576	0.0702	0.2750	0.9781
IPWP	0.0098	-0.0290	0.0825	0.0527	0.2064	0.7865	0.0106	-0.0324	0.0766	0.0592	0.2319	0.7639
rIPWP	0.0085	-0.0299	0.0771	0.0582	0.2282	0.7796	0.0101	-0.0344	0.0737	0.0585	0.2294	0.7456
IPWSy	0.0032	0.0018	0.0439	0.0446	0.1750	0.8797	0.0033	0.0015	0.0458	0.0448	0.1755	0.8726
rIPWSy	0.0033	0.0011	0.0440	0.0446	0.1748	0.8769	0.0033	0.0009	0.0458	0.0447	0.1752	0.8689
rIPWSry	0.0032	0.0006	0.0456	0.0461	0.1808	0.8896	0.0032	0.0002	0.0482	0.0471	0.1847	0.8901
IPWSt	0.0029	0.0016	0.0479	0.0496	0.1946	0.9184	0.0030	0.0008	0.0509	0.0506	0.1985	0.9204
rIPWSt	0.0029	0.0010	0.0482	0.0495	0.1942	0.9161	0.0031	0.0002	0.0508	0.0505	0.1981	0.9185
OutlierRatio=0.2												
IPWN	0.3249	-0.0396	0.5228	0.6136	2.4053	0.9690	0.3344	-0.0118	0.5080	0.7773	3.0468	0.9930
rIPWN	0.0134	0.0077	0.0851	0.0964	0.3778	0.9075	0.0120	0.0057	0.1028	0.1215	0.4764	0.9762
IPWP	0.4125	-0.0727	0.5962	0.6097	2.3900	0.9378	0.5221	-0.0155	0.6691	0.7681	3.0109	0.9626
rIPWP	0.0281	-0.0258	0.1201	0.0910	0.3566	0.7622	0.0378	-0.0261	0.1316	0.1046	0.4100	0.7644
IPWSy	0.3245	-0.0406	0.5198	0.5590	2.1910	0.9426	0.3396	-0.0156	0.5079	0.5597	2.1939	0.9325
rIPWSy	0.0138	0.0052	0.1151	0.1150	0.4507	0.9494	0.0143	-0.0007	0.1182	0.1183	0.4637	0.9559
rIPWSry	0.0144	-0.0016	0.1184	0.1175	0.4605	0.9526	0.0149	0.0002	0.1208	0.1215	0.4762	0.9608
IPWSt	0.3284	-0.0466	0.5232	0.5543	2.1727	0.9443	0.3465	-0.0211	0.5145	0.5313	2.0825	0.9128
rIPWSt	0.0141	-0.0013	0.1166	0.1132	0.4436	0.9470	0.0144	-0.0010	0.1173	0.1146	0.4491	0.9483
OutlierRatio=0.4												
IPWN	0.6268	0.0395	0.7406	0.8419	3.3001	0.9674	0.5970	-0.0187	0.7196	1.1173	4.3799	0.9976
rIPWN	0.0494	0.0043	0.2052	0.2166	0.8493	0.9538	0.0513	0.0277	0.1994	0.2434	0.9542	0.9710
IPWP	0.7628	0.0175	0.8271	0.8432	3.3052	0.9478	0.9809	-0.0555	0.9545	1.1244	4.4076	0.9783
rIPWP	0.0678	-0.0226	0.1974	0.1830	0.7174	0.8241	0.0906	-0.0397	0.2058	0.2166	0.8491	0.8246
IPWSy	0.6411	0.0375	0.7605	0.7780	3.0496	0.9525	0.6626	-0.0201	0.7732	0.8143	3.1920	0.9561
rIPWSy	0.0495	0.0015	0.2021	0.2073	0.8125	0.9325	0.0489	0.0078	0.2043	0.2112	0.8280	0.9422
rIPWSry	0.0497	0.0015	0.2031	0.2091	0.8195	0.9340	0.0484	0.0076	0.2034	0.2107	0.8258	0.9400
IPWSt	0.6445	0.0375	0.7629	0.7792	3.0542	0.9500	0.6833	-0.0127	0.7752	0.7411	2.9052	0.9098
rIPWSt	0.0502	-0.0005	0.2019	0.2040	0.7997	0.9259	0.0491	0.0071	0.2005	0.2002	0.7847	0.9221

Table 4.2. The distribution of $\hat{\tau}(x_1)$ for Model 2.

	d=4						d=10					
Method	MISE	Bias	SD	SE	Width	Rate	MISE	Bias	SD	SE	Width	Rate
OutlierRatio=0												
IPWN	0.0230	0.0532	0.1281	0.1514	0.5935	0.9485	0.0283	0.0611	0.1321	0.1720	0.6743	0.9577
rIPWN	0.0232	0.0537	0.1287	0.1521	0.5964	0.9489	0.0283	0.0609	0.1324	0.1725	0.6760	0.9577
IPWP	0.0207	0.0149	0.1268	0.1571	0.6159	0.9683	0.0239	0.0183	0.1216	0.1833	0.7184	0.9831
rIPWP	0.0209	0.0151	0.1273	0.1578	0.6186	0.9683	0.0239	0.0180	0.1220	0.1831	0.7177	0.9832
IPWSy	0.0203	0.0235	0.1182	0.1468	0.5756	0.9601	0.0223	0.0119	0.1463	0.1628	0.6382	0.9596
rIPWSy	0.0204	0.0239	0.1187	0.1476	0.5785	0.9610	0.0224	0.0118	0.1467	0.1639	0.6424	0.9602
rIPWSry	0.0230	0.0416	0.1103	0.1248	0.4894	0.9111	0.0234	0.0366	0.1140	0.1263	0.4950	0.9106
IPWSt	0.0225	0.0271	0.1201	0.1371	0.5376	0.9341	0.0224	0.0296	0.1327	0.1571	0.6159	0.9601
rIPWSt	0.0227	0.0274	0.1207	0.1379	0.5404	0.9337	0.0226	0.0297	0.1334	0.1578	0.6184	0.9598
OutlierRatio=0.2												
IPWN	0.5479	0.0261	0.6887	0.7575	2.9693	0.9622	0.5329	0.0667	0.6768	0.9807	3.8441	0.9962
rIPWN	0.0696	0.0608	0.2141	0.2450	0.9602	0.9399	0.0709	0.0598	0.2412	0.3151	1.2350	0.9817
IPWP	0.5497	-0.0714	0.6879	0.7569	2.9672	0.9624	0.5398	-0.0292	0.6868	0.9906	3.8831	0.9937
rIPWP	0.0629	0.0258	0.2108	0.2549	0.9990	0.9591	0.0662	0.0245	0.2135	0.3124	1.2246	0.9837
IPWSy	0.5470	-0.0070	0.6877	0.6555	2.5697	0.9235	0.5418	0.0369	0.6806	0.6900	2.7047	0.9355
rIPWSy	0.0699	0.0407	0.2491	0.2571	1.0077	0.9458	0.0714	0.0620	0.2165	0.2336	0.9157	0.9196
rIPWSry	0.0657	0.0337	0.2347	0.2345	0.9193	0.9298	0.0649	0.0472	0.2317	0.2369	0.9286	0.9440
IPWSt	0.5452	-0.0033	0.6868	0.6616	2.5934	0.9235	0.5259	0.0424	0.6691	0.7434	2.9140	0.9606
rIPWSt	0.0696	0.0319	0.2474	0.2622	1.0278	0.9522	0.0690	0.0549	0.2198	0.2506	0.9825	0.9431
OutlierRatio=0.4												
IPWN	0.9808	0.0265	0.9537	1.1054	4.3329	0.9694	0.9795	0.0650	0.9484	1.4000	5.4879	0.9930
rIPWN	0.1896	0.0991	0.3766	0.4382	1.7176	0.9516	0.1857	0.0822	0.3329	0.5243	2.0552	0.9902
IPWP	1.0072	-0.0222	0.9687	1.1185	4.3845	0.9674	1.0382	-0.0404	0.9808	1.4268	5.5928	0.9922
rIPWP	0.1780	0.0272	0.3767	0.4443	1.7417	0.9599	0.1846	-0.0078	0.3822	0.5598	2.1943	0.9903
IPWSy	1.0328	0.0553	0.9737	0.9476	3.7147	0.9355	0.9935	0.0248	0.9551	1.0090	3.9552	0.9368
rIPWSy	0.1835	0.0641	0.3724	0.3999	1.5676	0.9359	0.1852	0.0487	0.3784	0.4241	1.6626	0.9522
rIPWSry	0.1840	0.0822	0.3647	0.3558	1.3945	0.8934	0.1854	0.0680	0.3680	0.3571	1.3996	0.9019
IPWSt	1.0322	0.0612	0.9732	0.9420	3.6927	0.9297	0.9817	0.0311	0.9487	1.0559	4.1392	0.9605
rIPWSt	0.1810	0.0513	0.3702	0.3979	1.5596	0.9259	0.1824	0.0424	0.3773	0.4464	1.7499	0.9675

Table 4.3. The distribution of $\hat{\tau}(x_1)$ for Model 3.

	d=4						d=10					
Method	MISE	Bias	SD	SE	Width	Rate	MISE	Bias	SD	SE	Width	Rate
OutlierRatio=0												
IPWN	0.0150	0.0129	0.0901	0.1006	0.3944	0.9035	0.0137	0.0171	0.1067	0.1289	0.5051	0.9719
rIPWN	0.0150	0.0129	0.0905	0.1007	0.3947	0.9029	0.0138	0.0172	0.1069	0.1287	0.5044	0.9719
IPWP	0.0149	-0.0011	0.0910	0.1014	0.3976	0.9029	0.0144	-0.0084	0.1123	0.1299	0.5093	0.9616
rIPWP	0.0150	-0.0010	0.0914	0.1015	0.3980	0.9031	0.0145	-0.0085	0.1126	0.1297	0.5085	0.9615
IPWSy	0.0149	0.0091	0.1208	0.1296	0.5079	0.9591	0.0134	0.0094	0.1078	0.1116	0.4376	0.9435
rIPWSy	0.0150	0.0091	0.1211	0.1300	0.5095	0.9589	0.0135	0.0094	0.1082	0.1120	0.4391	0.9441
rIPWSry	0.0137	0.0162	0.1102	0.1110	0.4353	0.9308	0.0139	0.0222	0.1091	0.1110	0.4352	0.9335
IPWSt	0.0142	0.0128	0.1156	0.1191	0.4668	0.9435	0.0156	0.0195	0.1218	0.1291	0.5061	0.9542
rIPWSt	0.0143	0.0129	0.1159	0.1195	0.4685	0.9438	0.0157	0.0196	0.1220	0.1295	0.5076	0.9549
OutlierRatio=0.2												
IPWN	0.3509	0.0314	0.5227	0.6083	2.3847	0.9550	0.3542	-0.0049	0.5423	0.8085	3.1693	0.9915
rIPWN	0.0404	0.0274	0.1807	0.1925	0.7544	0.9457	0.0416	0.0261	0.1740	0.2146	0.8414	0.9638
IPWP	0.3491	0.0034	0.5221	0.6079	2.3829	0.9568	0.3657	0.0141	0.5467	0.7859	3.0806	0.9915
rIPWP	0.0393	0.0053	0.1797	0.1928	0.7559	0.9485	0.0430	0.0000	0.1814	0.2164	0.8482	0.9612
IPWSy	0.3493	0.0251	0.5207	0.5307	2.0803	0.9116	0.3601	-0.0299	0.5370	0.5679	2.2263	0.9255
rIPWSy	0.0409	0.0094	0.1753	0.1755	0.6879	0.9083	0.0402	0.0170	0.1764	0.1841	0.7217	0.9292
rIPWSry	0.0410	0.0211	0.1662	0.1606	0.6296	0.8846	0.0406	0.0179	0.1629	0.1601	0.6276	0.8839
IPWSt	0.3500	0.0254	0.5215	0.5315	2.0835	0.9144	0.3616	-0.0212	0.5399	0.5994	2.3497	0.9389
rIPWSt	0.0406	0.0078	0.1739	0.1754	0.6875	0.9165	0.0400	0.0097	0.1783	0.1884	0.7387	0.9340
OutlierRatio=0.4												
IPWN	0.6796	0.0201	0.7795	0.8757	3.4325	0.9596	0.6602	0.0524	0.7671	1.1549	4.5272	0.9956
rIPWN	0.1105	0.0133	0.2612	0.3068	1.2026	0.9317	0.1158	0.0086	0.2723	0.3922	1.5375	0.9763
IPWP	0.6784	0.0040	0.7792	0.8785	3.4437	0.9620	0.6718	0.0143	0.7763	1.1597	4.5460	0.9923
rIPWP	0.1115	-0.0030	0.2643	0.3075	1.2052	0.9298	0.1182	0.0023	0.3065	0.3952	1.5491	0.9772
IPWSy	0.6875	-0.0435	0.7785	0.7711	3.0228	0.9263	0.6414	0.0412	0.7508	0.8473	3.3215	0.9531
rIPWSy	0.1103	0.0193	0.2868	0.3031	1.1881	0.9241	0.1124	0.0156	0.2976	0.3241	1.2703	0.9429
rIPWSry	0.1098	0.0195	0.2796	0.2739	1.0737	0.8971	0.1105	0.0156	0.2831	0.2743	1.0753	0.9005
IPWSt	0.6971	0.0014	0.7905	0.7679	3.0101	0.9197	0.6423	0.0423	0.7516	0.8598	3.3702	0.9595
rIPWSt	0.1102	0.0177	0.2854	0.2972	1.1652	0.9210	0.1172	0.0107	0.3072	0.3312	1.2983	0.9417

2 and 3. This outcome is expected for two reasons: first, under Model 1, no dimension reduction occurs, so IPWSy and IPWN are theoretically identical, except for the estimation error of the rMAVE method. Second, since rMAVE also relies on kernel smoothing, its estimation error increases with dimension, so the smaller the dimension reduced by rMAVE, the less estimation error there will be. Therefore, it is reasonable that IPWSy has higher MISE than IPWN in Model 1.

For $d = 10$, both IPWSy and IPWSt successfully reduce the dimension to at least 4, and both methods have smaller MISE than IPWN.

4. Robust dimension reduction in contaminated cases.

In contaminated cases, rIPWSry slightly outperforms rIPWSy in Models 2 and 3. This is understandable, as robust rMAVE can provide more accurate results than rMAVE when the response is contaminated. For Model 1, a possible explanation is that rMAVE may suffer more from high dimensionality than rMAVE, leading to higher MISE. When the outlier ratio is high, rIPWSt outperforms rIPWSry, especially in Model 1, as T is not contaminated and can be reduced to lower dimensions, particularly in Models 1 and 3.

5. Confidence intervals.

For cases with $d = 10$, methods without dimension reduction such as IPWN, rIPWN, IPWP, and rIPWP tend to have much larger estimated standard errors than the true standard deviation, leading to overly conservative confidence intervals. This occurs because the sample size is insufficient to support the application of higher-order kernels required for the theoretical results to hold, and the estimation of the propensity score tends to overfit the data when the dimensionality is high. This causes propensity scores to frequently approach 0 or 1, making the variance estimator unstable. Simply increasing the bandwidth for estimating the propensity score does not significantly improve the variance estimator. However, this problem is substantially mitigated when dimension reduction is applied.

With dimension reduction, the widths of the confidence intervals generally align with the MISE comparison. Regarding coverage rates, IPWSy, rIPWSy, and rIPWSry achieve coverage rates close to the nominal 95% level in Models 2 and 3. IPWSt and rIPWSt achieve coverage rates close to the 95% nominal level only in Models 1 and 3. This outcome

can again be explained by the dimension that rMAVE ultimately reduces to, with greater estimation error and reduced coverage rates when the dimension is larger.

6. Estimation bias of kernel smoothing.

Even though higher-order kernels reduce the estimation bias, the bias still slightly affects the coverage rates of confidence intervals. This is particularly noticeable in cases without outliers, where the variance is smaller and bias becomes a larger issue compared to contaminated cases.

4.6 Real Data Analysis

In this section, we apply our proposed method to analyze data from the 2007-2008 NHANES , which assesses the health and nutrition statuses of both children and adults in the United States. Our objective is to examine whether participation in the National School Lunch Program (NSLP) affects Body Mass Index (BMI). In a related study, Huang and Chan (2017) estimated the CATE conditioned on a linear combination of all covariates using SDR. However, their method does not effectively capture the heterogeneity of treatment effects within subpopulations defined by a subset of covariates, providing an opportunity for us to demonstrate our proposed estimators. In this data application, we estimate the CATE conditioned solely on children's age to showcase our approach. All the estimators used in our simulations will be applied to highlight the differences.

4.6.1 The Dataset

Following the setup in Huang and Chan (2017), we include several covariates in our analysis: child age, gender, race, ratio of family income to poverty, WIC benefit (received in the last 12 months), child food security category, health insurance coverage, respondent's gender, and respondent's age. Child age is used as the primary conditioning variable for our CATE analysis. After creating dummy variables for the categorical covariates, we end up with a total of 11 covariates, making it essential to apply dimension reduction techniques. The response variable in our analysis is BMI and the treatment indicator is set by whether a child consumes school lunch more than two times per week. After removing missing values,

the dataset consists of 1049 children in the group having school meals and 241 in the group not having school meals.

To avoid the boundary effect, we estimate the CATE for children aged 6 to 15 years using all the methods mentioned in our simulations. All bandwidths are selected by cross-validation. The CATE estimators and the corresponding 95% confidence intervals are presented in the figures.

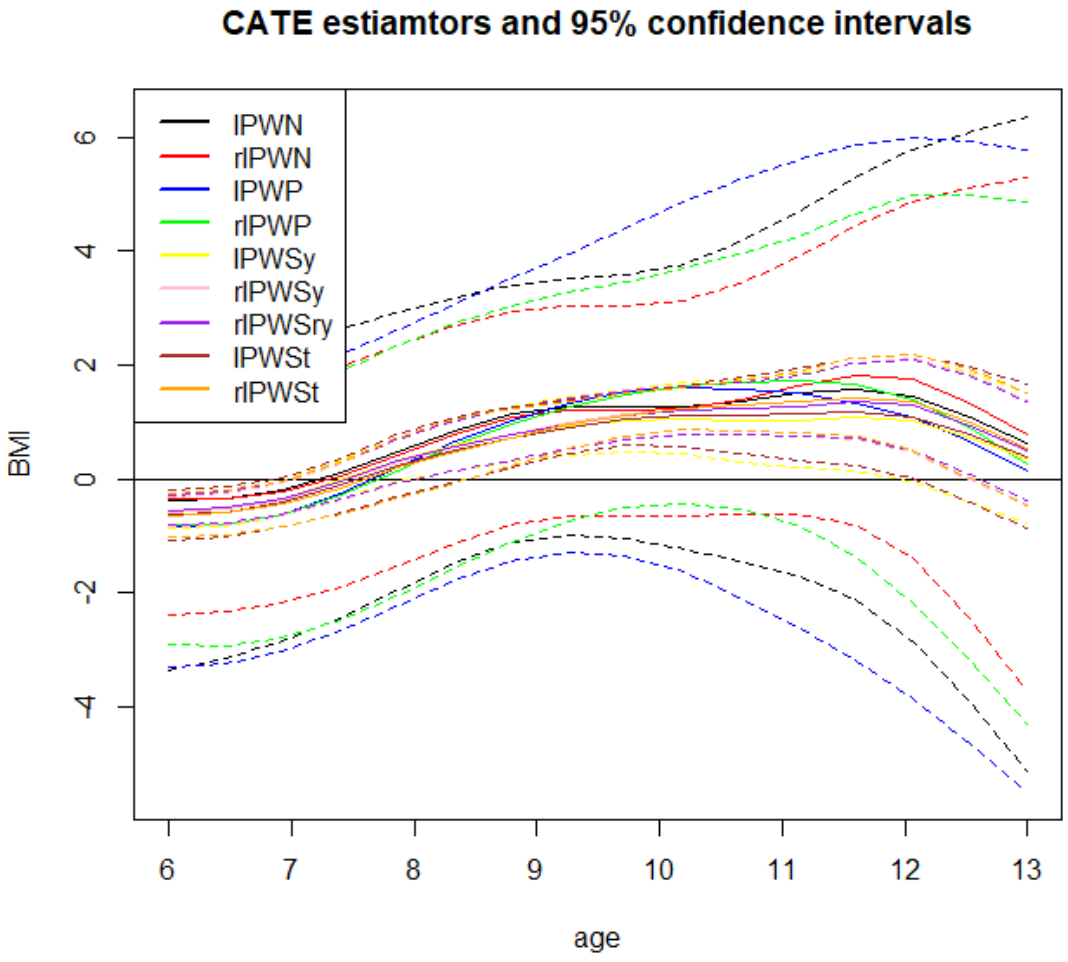

Figure 4.2. Comparison of all CATE estimators with 95% pointwise confidence bands in a single plot. Allowing direct visual comparison of their behavior across the children's age.

4.6.2 Estimation Result

Our analysis reveals that the overall trend in CATE estimates across various methods shows a similar pattern. The CATE increases at younger ages and then declines in older ages. Specifically, we observe a negative effect on BMI for children younger than 7 years, which turns positive for children older than 8. This result indicates that participating in the school lunch program may reduce BMI for younger children but increase it as they grow older. The CATE estimators of IPWP and rIPWP differ slightly from the others, as they are smoother and start decreasing earlier. However, we cannot conclude whether this is due to model misspecification or if logistic regression tends to provide smoother estimators.

The most notable differences across methods are observed in the confidence intervals. Consistent with our simulation results, the confidence intervals for IPWN, rIPWN, IPWP, and rIPWP tend to be significantly wider, likely reflecting the overfitting of the propensity score in these methods. In contrast, the robust methods, excluding those involving dimension reduction, yield narrower confidence intervals, demonstrating their effectiveness in handling heavy-tailed error distributions. Among the methods, rIPWSry exhibits the smallest confidence interval widths across age groups. Based on our conclusions from the simulation section, we choose this estimator as the proposed one for this data analysis.

4.6.3 Interpretation

In contrast to the findings in Huang and Chan (2017), where the treatment effect was found to be non-significant, our results show significant heterogeneity in CATE across different age groups. We find with 95% confidence that the CATE is significantly negative for children younger than 7 and significantly positive for children between the ages of 8 and 13.

These findings align with some existing literature. For instance, Winpenny et al. (2017) reports that children aged 10 to 14 who participate in school lunch programs tend to consume fewer fruits and vegetables and more fried foods and sugar-sweetened beverages during school hours. On the other hand, younger children in kindergarten appear to have more nutritious and lower-calorie school meals compared to packed lunches (Farris et al., 2014). These

behavioral differences may explain the age-dependent effects of school lunch participation on BMI found in our study.

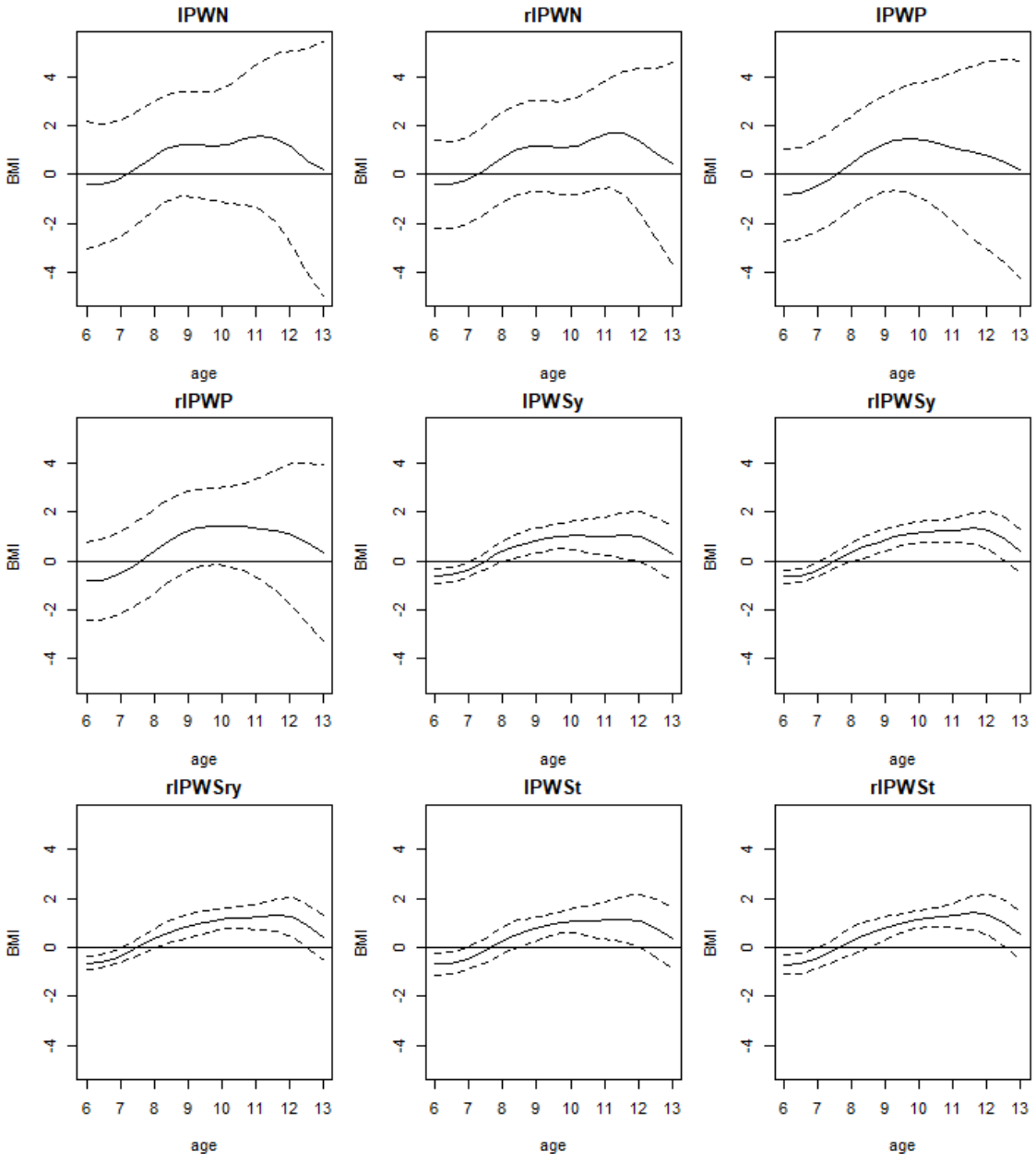

Figure 4.3. Individual plots of CATE estimators over children's age with 95% pointwise confidence bands.

REFERENCES

[1] Jersey Neyman. "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes". In: *Roczniki Nauk Rolniczych* 10.1 (1923), pp. 1–51.

[2] George EP Box and David R Cox. "An analysis of transformations". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 26.2 (1964), pp. 211–243.

[3] Peter J Huber. "Robust Estimation of a Location Parameter". In: *Ann. Math. Statist.* 35.4 (1964), pp. 73–101.

[4] Elizbar A Nadaraya. "On estimating regression". In: *Theory of Probability & Its Applications* 9.1 (1964), pp. 141–142.

[5] Geoffrey S Watson. "Smooth regression analysis". In: *Sankhy: The Indian Journal of Statistics, Series A* (1964), pp. 359–372.

[6] R Raj Bahadur. "A note on quantiles in large samples". In: *The Annals of Mathematical Statistics* 37.3 (1966), pp. 577–580.

[7] Peter J Huber. "Robust regression: asymptotics, conjectures and Monte Carlo". In: *The annals of statistics* (1973), pp. 799–821.

[8] Albert E Beaton and John W Tukey. "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data". In: *Technometrics* 16.2 (1974), pp. 147–185.

[9] Frank R Hampel. "The influence curve and its role in robust estimation". In: *Journal of the american statistical association* 69.346 (1974), pp. 383–393.

[10] Donald B Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[11] Paul W Holland and Roy E Welsch. "Robust regression using iteratively reweighted least-squares". In: *Communications in Statistics-theory and Methods* 6.9 (1977), pp. 813–827.

[12] Thomas F Moberg, John S Ramberg, and Ronald H Randles. "An adaptive M-estimator and its application to a selection problem". In: *Technometrics* 20.3 (1978), pp. 255–263.

[13] Peter J Bickel and Kjell A Doksum. "An analysis of transformations revisited". In: *Journal of the american statistical association* 76.374 (1981), pp. 296–311.

[14] David L Donoho and Peter J Huber. "The notion of breakdown point". In: *A festschrift for Erich L. Lehmann* 157184 (1983).

[15] Naihua Duan. "Smearing estimate: a nonparametric retransformation method". In: *Journal of the American Statistical Association* 78.383 (1983), pp. 605–610.

[16] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

[17] Wolfgang Härdle. "Robust regression function estimation". In: *Journal of Multivariate Analysis* 14.2 (1984), pp. 169–180.

[18] Peter J Huber. "Finite Sample Breakdown of M-and P-Estimators". In: *The Annals of Statistics* 12.1 (1984), pp. 119–126.

[19] Don M Miller. "Reducing transformation bias in curve fitting". In: *The American Statistician* 38.2 (1984), pp. 124–126.

[20] WOLFGANG Hardle. "A note on jackknifing kernel regression function estimators (corresp.)" In: *IEEE transactions on information theory* 32.2 (1986), pp. 298–300.

[21] Peter J Rousseeuw et al. *Robust statistics: the approach based on influence functions*. 1986.

[22] Arthur S Leon et al. "Leisure-time physical activity levels and risk of coronary heart disease and death: the Multiple Risk Factor Intervention Trial". In: *Jama* 258.17 (1987), pp. 2388–2395.

[23] J Steve Marron and Matt P Wand. "Exact mean integrated squared error". In: *The Annals of Statistics* 20.2 (1992), pp. 712–736.

[24] JS Marron. "Visual understanding of higher-order kernels". In: *Journal of Computational and Graphical Statistics* 3.4 (1994), pp. 447–458.

[25] Morten Mowé and Thomas Bøhmer. "Increased levels of alcohol markers (γGT, MCV, ASAT, ALAT) in older patients are not related to high alcohol intake". In: *Journal of the American Geriatrics Society* 44.9 (1996), pp. 1136–1137.

[26] Jinyong Hahn. "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". In: *Econometrica* (1998), pp. 315–331.

[27] Rainer Hambrecht, Eduard Fiehn, et al. "Regular physical exercise corrects endothelial dysfunction and improves exercise capacity in patients with chronic heart failure". In: *Circulation* 98.24 (1998), pp. 2709–2715.

[28] Guoying Li and Jian Zhang. In: *The Annals of Statistics* 26.3 (1998), pp. 1170–1189.

[29] Edoardo Giannini et al. "Progressive liver functional impairment is associated with an increase in AST/ALT ratio". In: *Digestive diseases and sciences* 44 (1999), pp. 1249–1253.

[30] Adrian Pagan et al. *Nonparametric econometrics*. Vol. 10. Cambridge university press Cambridge, 1999.

[31] Rainer Hambrecht, Anamaria Wolf, et al. "Effect of exercise on coronary endothelial function in patients with coronary artery disease". In: *New England Journal of Medicine* 342.7 (2000), pp. 454–460.

[32] United States Department of Health and Human Services. *Healthy People 2010*. 2nd. Washington, DC: US Government Printing Office, 2000.

[33] D Wang, JA Romagnoli, and AA Safavi. "Wavelet-based adaptive robust M-estimator for nonlinear system identification". In: *AIChE Journal* 46.8 (2000), pp. 1607–1615.

[34] R Dennis Cook and Bing Li. "Dimension reduction for conditional mean in regression". In: *The Annals of Statistics* 30.2 (2002), pp. 455–474.

[35] Ichiro Takeuchi, Yoshua Bengio, and Takafumi Kanamori. "Robust regression with asymmetric heavy-tail noise distributions". In: *Neural Computation* 14.10 (2002), pp. 2469–2496.

[36] Yingcun Xia et al. "An adaptive estimation of dimension reduction space". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 363–410.

[37] Katherine M Conigrave et al. “Traditional markers of excessive alcohol use”. In: *Addiction* 98 (2003), pp. 31–43.

[38] Keisuke Hirano, Guido W Imbens, and Geert Ridder. “Efficient estimation of average treatment effects using the estimated propensity score”. In: *Econometrica* 71.4 (2003), pp. 1161–1189.

[39] Jinyong Hahn. “Functional restriction and efficiency in causal inference”. In: *The Review of Economics and Statistics* 86.1 (2004), pp. 73–76.

[40] Guido W Imbens. “Nonparametric estimation of average treatment effects under exogeneity: A review”. In: *Review of Economics and statistics* 86.1 (2004), pp. 4–29.

[41] National Institute of Alcohol Abuse and Alcoholism (NIAAA). “NIAAA council approves definition of binge drinking”. In: *NIAAA Newsletter* 3.3 (2004).

[42] Council of State et al. “Indicators for chronic disease surveillance”. In: *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and Reports* 53.RR-11 (2004), pp. 1–6.

[43] United States Department of Agriculture, United States Department of Health, and Human Services. *Dietary Guidelines for Americans*. Washington, DC: US Government Printing Office, 2005. Chap. Chapter 9. Alcoholic Beverages, pp. 43–46.

[44] Wayne D Comper and Tanya M Osicka. “Detection of urinary albumin”. In: *Advances in chronic kidney disease* 12.2 (2005), pp. 170–176.

[45] Avi Giloni and Jeffrey S Simonoff. “The conditional breakdown properties of least absolute value local polynomial estimators”. In: *Journal of Nonparametric Statistics* 17.1 (2005), pp. 15–30.

[46] Johanna Hietala et al. “Serum gamma-glutamyl transferase in alcoholics, moderate drinkers and abstainers: effect on gt reference intervals at population level”. In: *Alcohol and Alcoholism* 40.6 (2005), pp. 511–514.

[47] M Alan Brookhart et al. “Variable selection for propensity score models”. In: *American journal of epidemiology* 163.12 (2006), pp. 1149–1156.

[48] Pavel íek and Wolfgang Härdle. “Robust estimation of dimension reduction space”. In: *Computational statistics & data analysis* 51.2 (2006), pp. 545–555.

[49] Takafumi Kanamori and Ichiro Takeuchi. “Conditional mean estimation under asymmetric and heteroscedastic error by linear combination of quantile regressions”. In: *Computational statistics & data analysis* 50.12 (2006), pp. 3605–3618.

[50] Mandy Stahre et al. “Measuring average alcohol consumption: the impact of including binge drinks in quantity–frequency calculations”. In: *Addiction* 101.12 (2006), pp. 1711–1718.

[51] Sergio Firpo. “Efficient semiparametric estimation of quantile treatment effects”. In: *Econometrica* 75.1 (2007), pp. 259–276.

[52] Yingcun Xia. “A constructive approach to the estimation of dimension reduction directions”. In: (2007).

[53] Marit D Solbu et al. “Cardiovascular risk-factors predict progression of urinary albumin-excretion in a general, non-diabetic population: a gender-specific follow-up study”. In: *Atherosclerosis* 201.2 (2008), pp. 398–406.

[54] Päivikki Alatalo et al. “Biomarkers of liver status in heavy drinkers, moderate drinkers and abstainers”. In: *Alcohol & Alcoholism* 44.2 (2009), pp. 199–203.

[55] Guido W Imbens and Geert Ridder. “Estimation and Inference for generalized full and partial means and average derivatives”. In: *URL http://www. economics. harvard. edu/faculty/imbens/files/ir. pdf* (2009).

[56] W Greg Miller et al. “Current issues in measurement and reporting of urinary albumin excretion”. In: *Clinical chemistry* 55.1 (2009), pp. 24–38.

[57] Hugh A Chipman, Edward I George, and Robert E McCulloch. “BART: Bayesian additive regression trees”. In: (2010).

[58] Adam N Glynn and Kevin M Quinn. “An introduction to the augmented inverse propensity weighted estimator”. In: *Political analysis* 18.1 (2010), pp. 36–56.

[59] Emily S Robinson et al. “Physical activity and albuminuria”. In: *American journal of epidemiology* 171.5 (2010), pp. 515–521.

[60] Xavier De Luna, Ingeborg Waernbaum, and Thomas S Richardson. “Covariate selection for the nonparametric estimation of an average treatment effect”. In: *Biometrika* 98.4 (2011), pp. 861–875.

[61] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. "Subgroup identification from randomized clinical trial data". In: *Statistics in medicine* 30.24 (2011), pp. 2867–2880.

[62] Jennifer L Hill. "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 217–240.

[63] Pranab K Sen. *Introduction to nonparametric estimation by Alexandre B. Tsybakov.* 2011.

[64] Olivier Catoni. "Challenging the empirical mean and empirical variance: a deviation study". In: *Annales de l'IHP Probabilités et statistiques.* Vol. 48. 4. 2012, pp. 1148–1185.

[65] Yanyuan Ma and Liping Zhu. "A semiparametric approach to dimension reduction". In: *Journal of the American Statistical Association* 107.497 (2012), pp. 168–179.

[66] Zhiwei Zhang et al. "Causal inference on quantiles with an obstetric application". In: *Biometrics* 68.3 (2012), pp. 697–706.

[67] Alex Chang et al. "Lifestyle-related factors, obesity, and incident microalbuminuria: the CARDIA (Coronary Artery Risk Development in Young Adults) study". In: *American journal of kidney diseases* 62.2 (2013), pp. 267–275.

[68] Weixin Yao and Qin Wang. "Robust variable selection through MAVE". In: *Computational Statistics & Data Analysis* 63 (2013), pp. 42–49.

[69] Alisha R Farris et al. "Nutritional comparison of packed and school lunches in pre-kindergarten and kindergarten children following the implementation of the 2012–2013 national school lunch program standards". In: *Journal of nutrition education and behavior* 46.6 (2014), pp. 621–626.

[70] Prasad P Torkadi, IC Apte, and AK Bhute. "Biochemical evaluation of patients of alcoholic liver disease and non-alcoholic liver disease". In: *Indian journal of clinical biochemistry* 29 (2014), pp. 79–83.

[71] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. "Estimating conditional average treatment effects". In: *Journal of Business & Economic Statistics* 33.4 (2015), pp. 485–505.

[72] Sanjiv Agarwal, Victor L Fulgoni, and Harris R Lieberman. "Assessing alcohol intake & its dose-dependent effects on liver enzymes by 24-h recall and questionnaire using NHANES 2001-2010 data". In: *Nutrition Journal* 15 (2015), pp. 1–12.

[73] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7353–7360.

[74] Jianqing Fan, Quefeng Li, and Yuyan Wang. "Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.1 (2017), pp. 247–265.

[75] Pierre Gutierrez and Jean-Yves Gérardy. "Causal inference and uplift modelling: A review of the literature". In: *International conference on predictive applications and APIs*. PMLR. 2017, pp. 1–13.

[76] Ming-Yueh Huang and Kwun Chuen Gary Chan. "Joint sufficient dimension reduction and estimation of conditional and average treatment effects". In: *Biometrika* 104.3 (2017), pp. 583–596.

[77] Ming-Yueh Huang and Chin-Tsang Chiang. "An effective semiparametric estimation approach for the sufficient dimension reduction model". In: *Journal of the American Statistical Association* 112.519 (2017), pp. 1296–1310.

[78] Sokbae Lee, Ryo Okui, and Yoon-Jae Whang. "Doubly robust uniform confidence band for the conditional average treatment effect function". In: *Journal of Applied Econometrics* 32.7 (2017), pp. 1207–1225.

[79] Wei Luo, Yeying Zhu, and Debashis Ghosh. "On estimating regression-based causal effects using sufficient dimension reduction". In: *Biometrika* 104.1 (2017), pp. 51–65.

[80] Susan M Shortreed and Ashkan Ertefaie. "Outcome-adaptive lasso: variable selection for causal inference". In: *Biometrics* 73.4 (2017), pp. 1111–1122.

[81] Eleanor M Winpenny et al. "Changes in diet from age 10 to 14 years and prospective associations with school lunch choice". In: *Appetite* 116 (2017), pp. 259–267.

[82] Joshua Lang et al. "Association of serum albumin levels with kidney function decline and incident chronic kidney disease in elders". In: *Nephrology Dialysis Transplantation* 33.6 (2018), pp. 986–992.

[83] Bing Li. *Sufficient dimension reduction: Methods and applications with R*. CRC Press, 2018.

[84] Scott Powers et al. "Some methods for heterogeneous treatment effect estimation in high dimensions". In: *Statistics in medicine* 37.11 (2018), pp. 1767–1787.

[85] Liuyi Yao et al. "Representation learning for treatment effect estimation from observational data". In: *Advances in neural information processing systems* 31 (2018).

[86] Wen-Xin Zhou et al. "A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing". In: *Annals of statistics* 46.5 (2018), p. 1904.

[87] Susan Athey, Julie Tibshirani, and Stefan Wager. "Generalized random forests". In: (2019).

[88] Pauline Burke et al. "Measuring Average Treatment Effect from Heavy-tailed Data". In: *arXiv preprint arXiv:1905.09252* (2019).

[89] Vincent Dorie et al. "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition". In: (2019).

[90] Sören R Künzel et al. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10 (2019), pp. 4156–4165.

[91] Fredrik Åberg, Martti Färkkilä, and Ville Männistö. "Interaction between alcohol use and metabolic risk factors for liver disease: a critical review of epidemiological studies". In: *Alcoholism: Clinical and Experimental Research* 44.2 (2020), pp. 384–403.

[92] Ming-Yueh Huang and Shu Yang. "Robust inference of conditional average treatment effects using dimension reduction". In: *arXiv preprint arXiv:2008.13137* (2020).

[93] Wei Luo and Yeying Zhu. "Matching using sufficient dimension reduction for causal inference". In: *Journal of Business & Economic Statistics* 38.4 (2020), pp. 888–900.

[94] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. "Adaptive huber regression". In: *Journal of the American Statistical Association* 115.529 (2020), pp. 254–265.

[95] Ying Zhang et al. “Quantile treatment effect estimation with dimension reduction”. In: *Statistical Theory and Related Fields* 4.2 (2020), pp. 202–213.

[96] Liya Fu and You-Gan Wang. “Robust regression with asymmetric loss functions”. In: *Statistical Methods in Medical Research* 30.8 (2021), pp. 1800–1815.

[97] Kazuharu Harada and Hironori Fujisawa. “Outlier-Resistant Estimators for Average Treatment Effect in Causal Inference”. In: *arXiv preprint arXiv:2106.13946* (2021).

[98] Xinkun Nie and Stefan Wager. “Quasi-oracle estimation of heterogeneous treatment effects”. In: *Biometrika* 108.2 (2021), pp. 299–319.

[99] Jing Zhang, Qin Wang, et al. “Robust MAVE through nonconvex penalized regression”. In: *Computational Statistics & Data Analysis* 160 (2021), p. 107247.

[100] Niwen Zhou and Lixing Zhu. “On IPW-based estimation of conditional average treatment effects”. In: *Journal of Statistical Planning and Inference* 215 (2021), pp. 1–22.

[101] Qingliang Fan et al. “Estimation of conditional average treatment effects with high-dimensional data”. In: *Journal of Business & Economic Statistics* 40.1 (2022), pp. 313–327.

[102] Hsin-Yi Kuo et al. “The Effects of Exercise Habit on Albuminuria and Metabolic Indices in Patients with Type 2 Diabetes Mellitus: A Cross-Sectional Study”. In: *Medicina* 58.5 (2022), p. 577.

[103] Christoph F Kurz. “Augmented inverse probability weighting and the double robustness property”. In: *Medical Decision Making* 42.2 (2022), pp. 156–167.

[104] Lu Li, Niwen Zhou, and Lixing Zhu. “Outcome regression-based estimation of conditional average treatment effect”. In: *Annals of the Institute of Statistical Mathematics* 74.5 (2022), pp. 987–1041.

[105] Jiyu Luo, Qiang Sun, and Wen-Xin Zhou. “Distributed adaptive Huber regression”. In: *Computational Statistics & Data Analysis* 169 (2022), p. 107419.

[106] Haoran Zhao et al. “Matching Using Sufficient Dimension Reduction for Heterogeneity Causal Effect Estimation”. In: *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*. IEEE. 2022, pp. 9–16.

[107] Vasimahmed Lala, Muhammad Zubair, and David Minter. “Liver function tests”. In: *StatPearls* (2023).

A. APPENDIX OF OUTLIER RESISTANT INFERENCE FOR CONDITIONAL AVERAGE TREATMENT EFFECT

A.1 The Proof of Theorem 2.2.1

Proof:

$$\begin{aligned}
E\left[\frac{T}{\pi(X)}\psi(Y-\mu_1(x_1)) \mid X_1=x_1\right] &= E\left[E\left[\frac{T}{\pi(X)}\psi(Y-\mu_1(x_1)) \mid X\right] \mid X_1=x_1\right] \\
&= E\left[E\left[\frac{T}{\pi(X)}\psi(Y^{(1)}-\mu_1(x_1)) \mid X\right] \mid X_1=x_1\right] \\
&= E\left[\frac{E[T \mid X]}{\pi(X)}E\left[\psi(Y^{(1)}-\mu_1(x_1)) \mid X\right] \mid X_1=x_1\right] \\
&= E\left[E\left[\psi(Y^{(1)}-\mu_1(x_1)) \mid X\right] \mid X_1=x_1\right] \\
&= E\left[\psi(Y^{(1)}-\mu_1(x_1)) \mid X_1=x_1\right] \\
&= \int \psi(y^{(1)}-\mu_1(x_1))f(y^{(1)} \mid x_1)\, dy^{(1)},
\end{aligned}$$

Since $\psi(y^{(1)}-\mu_1(x_1))$ is antisymmetric and $f(y^{(1)} \mid x_1)$ is symmetric with respect to $\mu_1(x_1)$, we have:

$$\int \psi(y^{(1)}-\mu_1(x_1))f(y^{(1)} \mid x_1)\, dy^{(1)} = 0.$$

.

A.2 The Proof of Theorem 2.3.1(1)

When $\sum_{y_i \in S_{Y_c}} W_i < A(\sum_{y_i \in S_{Y_c}} W_i + \sum_{y_i \in S_{Y_n}} W_i)$, there exits $\delta > 0$ s.t. $\sum_{y_i \in S_{Y_c}} W_i + \sum_{y_i \in S_{Y_n}} W_i\delta < A(\sum_{y_i \in S_{Y_n}} W_i + \sum_{y_i \in S_{Y_c}} W_i)$.

Let C be such that $\rho(x) \geq -\delta$ for $\mid x \mid \geq C$, let t be any real number such that $\mid y-t \mid \geq C$ for all $y \in Y$, then we have

$$\sum_{y_i \in S_{Y_n} \cup S_{Y_c}} \rho(y_i-\mu)W_{y_i} \leq -A\left(\sum_{y_i \in S_{Y_c}} W_i + \sum_{y_i \in S_{Y_n}} W_i\right),$$

$$\sum_{y_i \in S_{Y_n} \cup S_{Y_c}} \rho(y_i-t)W_{y_i} \geq -\delta \sum_{y_i \in S_{Y_n}} W_i - \sum_{y_i \in S_{Y_c}} W_i.$$

Hence we have

$$\sum_{y_i \in S_{Y_n} \cup S_{Y_c}} \rho(y_i - \mu) W_i < \sum_{y_i \in S_{Y_n} \cup S_{Y_c}} \rho(y_i - t) W_i.$$

Then μ must fall within distance C from a point in Y

When $\sum_{y_i \in S_{Y_c}} W_i > A(\sum_{y_i \in S_{Y_c}} W_i + \sum_{y_i \in S_{Y_n}} W_i)$.

Let $\delta > 0$ be such that $\sum_{y_i \in S_{Y_c}} W_i - \delta \sum_{y_i \in S_{Y_c}} W_i > A(\sum_{y_i \in S_{Y_c}} W_i + \sum_{y_i \in S_{Y_n}} W_i)$

Let C be such that $\rho(x) \geq -\delta$ for $\mid x \mid \geq C$.

Let k be any real number and assume that all points in S_{Y_c} are equal to k.

Assume all points in S_{Y_c} are equal to y^*. Then, for all t with $\mid y_i - t \mid \geq C$, we obtain

$$\sum_{y_i \in S_{Y_n} \cup S_{Y_c}} \rho(y_i - t) W_i \geq -A \left(\sum_{y_i \in S_{Y_c}} W_i + \sum_{y_i \in S_{Y_n}} W_i \right) - \delta \sum_{y_i \in S_{Y_c}} W_i$$

$$\sum_{y_i \in S_{Y_n} \cup S_{Y_c}} \rho(y_i - y^*) W_i \leq - \sum_{y_i \in S_{Y_c}} W_i$$

Then we have

$$\sum_{y_i \in S_{Y_n} \cup S_{Y_c}} \rho(y_i - y^*) W_i < \sum_{y_i \in S_{Y_n} \cup S_{Y_c}} \rho(y_i - t) W_i$$

Hence μ must in distance C from y^*. Let $y^* \to +\infty$ we have $\mu \to +\infty$.

A.3 The Proof of Theorem 2.3.1(2)

To prove the theorem, we can use lemma 4.2 in Huber (1984).

Let $M(t) = sup_x \mid \rho(x + t) - \rho(x) \mid$.

Since ρ is symmetric, we clearly have $M(-t) = M(t)$, and we may omit the absolute value bars in the definition without changing $M(t)$.

Lemma 4.2. The difference $\eta(t) = M(t) - \rho(t)$ is bounded:$0 \leq \eta(t) \leq x_0 \psi(x_0)$. For $t \geq x_0$, we have $\eta(t) \leq x_0 \psi(t)$, hence $\eta(t) \to 0$ for $t \to 0$.

And with lemma 4.2, we modified their lemma 4.3:

Modified Lamma 4.3. Put

$$\Delta_{Y\cup S_{Y_c}}(t) = \sum_{y_{\rm i}\in S_{Y_n}\cup S_{Y_c}} \rho(y_{\rm i}-t)-\rho(y_{\rm i})$$

Then there is a constant C which depends on Y and on $\sum\limits_{y_{\rm i}\in S_{Y_c}} W_{\rm i}$ but not on the actual values in S_{Y_c}, such that for all t

$$\rho(t)\left(\sum_{y_{\rm i}\in S_{Y_n}} W_{\rm i}-\sum_{y_{\rm i}\in S_{Y_c}} W_{\rm i}\right)-C\le\Delta_{Y\cup S_{Y_c}}(t)\le\rho(t)\left(\sum_{y_{\rm i}\in S_{Y_n}\cup S_{Y_c}} W_{\rm i}\right)+C$$

Proof:

$$\Delta_Y(t)=\sum_{y_{\rm i}\in S_{Y_n}}[\rho(y_{\rm i}-t)-\rho(y_{\rm i})]W_{\rm i}=\rho(t)\sum_{y_{\rm i}\in S_{Y_n}} W_{\rm i}+\sum_{y_{\rm i}\in S_{Y_n}}[\rho(y_{\rm i}-t)-\rho(t)]W_{\rm i}-\sum_{y_{\rm i}\in S_{Y_n}}\rho(y_{\rm i})W_{\rm i}.$$

Since $\mid \rho(y-t)-\rho(t) \mid = \mid \rho(t)-\rho(t-y) \mid \le \mid y \mid \psi(y_0)$, we have $\mid \Delta_Y(t)-\rho(t)\sum\limits_{y_{\rm i}\in S_{Y_n}} W_{\rm i} \mid \le C_1$ with $C_1=\sum\limits_{y_{\rm i}\in S_{Y_n}}\rho(y_{\rm i})W_{\rm i}+\sum\limits_{y_{\rm i}\in S_{Y_n}} \mid y_{\rm i} \mid \psi(y_0)W_{\rm i}$.

On the other hand

$$\mid \Delta_{S_{Y_c}}(t) \mid = \mid \sum_{y_{\rm i}\in S_{Y_c}}[\rho(y_{\rm i}-t)-\rho(y_{\rm i})]\,W_{\rm i} \mid \le M(t)\sum_{y_{\rm i}\in S_{Y_c}} W_{\rm i}=(\rho(t)+\eta(t))\sum_{y_{\rm i}\in S_{Y_c}} W_{\rm i}.$$

Since $\eta(t)$ is bounded $\mid \Delta_{S_{Y_c}}(t) \mid \le \rho(t)\sum_{y_{\rm i}\in S_{Y_c}} W_{\rm i}+C_2\sum_{y_{\rm i}\in S_{Y_c}} W_{\rm i}$.

Let$C=C_1+C_2\sum_{y_{\rm i}\in S_{Y_c}} W_{\rm i}$ we have

$$\rho(t)\left(\sum_{y_{\rm i}\in S_{Y_n}} W_{\rm i}-\sum_{y_{\rm i}\in S_{Y_c}} W_{\rm i}\right)-C\le\Delta_{Y\cup S_{Y_c}}(t)\le\rho(t)\left(\sum_{y_{\rm i}\in S_{Y_n}\cup S_{Y_c}} W_{\rm i}\right)+C.$$

With the modified lemma 4.3, we prove theorem 3.2.

Proof:

When

$$\sum_{y_{\rm i}\in S_{Y_c}} W_{\rm i}<\sum_{y_{\rm i}\in S_{Y_n}} W_{\rm i},$$

we have $\Delta_{Y\cup S_{Y_c}}(t)$ be bounded away from 0 for sufficiently large t, uniformly in Y.

Since $\Delta_{Y\cup S_{Y_c}}(0) = 0$ and since $\Delta_{Y\cup S_{Y_c}}(t)$ reaches its absolute minimum at μ, it follows that μ cannot be outside a certain bounded neighborhood of 0.

Hence we have μ bounded when

$$\sum_{y_i \in S_{Y_c}} W_i < \sum_{y_i \in S_{Y_n}} W_i.$$

Now we consider the opposite case, when $\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} < 1$.

Let $l(\mu) = \sum_{S_{Y_n}} \rho(y_i - \mu)W_i + \sum_{S_{Y_c}} \rho(y_i - \mu)W_i$ denote the loss function.

Let all the y_i in S_{Y_c} equals S_{Y_c} and solve the inequality $l(t) > l(S_{Y_c})$, that is

$$\sum_{S_{Y_n}} \rho(Y_i - t)W_i + \sum_{S_{Y_c}} \rho(Y_i - t)W_i > \sum_{S_{Y_n}} \rho(Y_i - S_{Y_c})W_i + \sum_{S_{Y_c}} \rho(Y_i - S_{Y_c})W_i,$$

we have

$$\frac{\rho(S_{Y_c} - t)}{\rho(y_i - S_{Y_c})} > \frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i}.$$

By property of ρ, we have $\frac{\rho(S_{Y_c}-t)}{\rho(y_i - S_{Y_c})} \to 1$ as $S_{Y_c} \to \infty$. Then for any fixed t the inequality holds under sufficiently large S_{Y_c}. Then we have for any fixed t, the loss function l does not attain its minimum at t.

Hence we have the estimator unbounded.

To completely show the breakdown property of μ under condition $2(U\rho)$ we need to show the estimator μ is going to be unbounded for some $y_i \in S_{Y_c}$ when

$$\sum_{y_i \in S_{Y_c}} W_i \geq \sum_{y_i \in S_{Y_n}} W_i.$$

To do this, we try to reverse the set of observations Y and S_{Y_c} by the following settings:

Let $y_i = S_{Y_c}$ for all $y_i \in S_{Y_c}$,

let $\tilde{y}_i = y_i - S_{Y_c} \in \tilde{Y}$ for $y_i \in S_{Y_n}$,

and let $\tilde{y}_i = y_i - S_{Y_c} \in \tilde{Y}^*$ for $y_i \in S_{Y_c}$.

For the estimator, we let $\tilde{\mu} = \mu - S_{Y_c}$.

When $\tilde{\mu}$ is bounded, we have $\mu \to \infty$ as $S_{Y_c} \to \infty$.

Then by the theorem 2, we have $\tilde{\mu}$ bounded for any choice of $\tilde{y}_\mathrm{i} \in \tilde{Y}$ when

$$\sum_{y_\mathrm{i} \in \tilde{Y}^*} W_\mathrm{i} > \sum_{y_\mathrm{i} \in \tilde{Y}} W_\mathrm{i}.$$

This is the same as we have μ unbounded for some choice of S_{Y_c} when

$$\sum_{y_\mathrm{i} \in S_{Y_c}} W_\mathrm{i} > \sum_{y_\mathrm{i} \in S_{Y_n}} W_\mathrm{i}.$$

A.4 The Proof of Theorem 2.3.1(3)

For the estimating equation:

$$0 = \sum_{y_\mathrm{i} \in S_{Y_n}} W_\mathrm{i} \psi(y_\mathrm{i} - \mu) + \sum_{y_\mathrm{i} \in S_{Y_c}} W_\mathrm{i} \psi(y_\mathrm{i} - \mu)$$

First, we assume μ is finite. Then for some C we have $\mid \mu \mid \leq C$ for all $y_\mathrm{i} \in S_{Y_c}$.

Hence we can take $y_\mathrm{i} = S_{Y_c}$ for $\forall y_\mathrm{i} \in S_{Y_c}$.

So that

$$0 = \sum_{y_\mathrm{i} \in S_{Y_n}} W_\mathrm{i} \psi(y_\mathrm{i} - \mu) + \sum_{y_\mathrm{i} \in S_{Y_c}} W_\mathrm{i} \psi(S_{Y_c} - \mu)$$

Let $S_{Y_c} \to \infty$. Since μ is bounded, we have $\psi(S_{Y_c} - \mu) \to k_2$.

And since $\psi \geq -k_1$, we have

$$0 \geq -k_1 \sum_{y_\mathrm{i} \in S_{Y_n}} W_\mathrm{i} + k_2 \sum_{y_\mathrm{i} \in S_{Y_c}} W_\mathrm{i},$$

$$k_1 \sum_{y_\mathrm{i} \in S_{Y_n}} W_\mathrm{i} \geq k_2 \sum_{y_\mathrm{i} \in S_{Y_c}} W_\mathrm{i}.$$

And let $y*^* \to -\infty$, we have $\psi(S_{Y_c} - \mu) = -K$ and since $\psi \leq k_2$, we have

$$0 \leq k_2 \sum_{y_\mathrm{i} \in S_{Y_n}} W_\mathrm{i} - k_1 \sum_{y_\mathrm{i} \in S_{Y_c}} W_\mathrm{i},$$

$$k_1 \sum_{y_i \in S_{Y_c}} W_i \leq k_2 \sum_{y_i \in S_{Y_n}} W_i.$$

Under condition 3, we have $k_1, k_2 > 0$ then we have: when μ is bounded, we have

$$\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} \geq \max\left\{\frac{k_2}{k_1}, \frac{k_1}{k_2}\right\}.$$

On the other hand, suppose there exists a sequence y_n such that μ_n is unbounded as a solution of

$$0 = \sum_{y_i \in S_{Y_n}} W_i \psi(y_i - \mu) + \sum_{y_i \in S_{Y_c}} W_i \psi(y_n - \mu).$$

Suppose there is a subsequence of $\{\mu_n\}$ tending to $+\infty$.

Then for this subsequence, $y_i - \mu_n \to -\infty$ for $y_i \in S_{Y_n}$, and since $\psi \leq k_2$ we have

$$0 \leq -k_1 \sum_{y_i \in S_{Y_n}} W_i + k_2 \sum_{y_i \in S_{Y_c}} W_i,$$

$$k_1 \sum_{y_i \in S_{Y_n}} W_i \leq k_2 \sum_{y_i \in S_{Y_c}} W_i.$$

Or if the subsequence $\mu_n \to +\infty$ for $y_i \in S_{Y_n}$, and since $\psi \geq k_1$ we have

$$0 \geq k_2 \sum_{y_i \in S_{Y_n}} W_i - k_1 \sum_{y_i \in S_{Y_c}} W_i,$$

$$k_1 \sum_{y_i \in S_{Y_c}} W_i \geq k_2 \sum_{y_i \in S_{Y_n}} W_i.$$

Then we have

$$\frac{\sum\limits_{y_i \in S_{Y_n}} W_i}{\sum\limits_{y_i \in S_{Y_c}} W_i} \leq \frac{k_1}{k_2},$$

or

$$\frac{\sum\limits_{y_i \in S_{Y_n}} W_i}{\sum\limits_{y_i \in S_{Y_c}} W_i} \leq \frac{k_2}{k_1}.$$

That is

$$\frac{\sum_{y_i \in S_{Y_n}} W_i}{\sum_{y_i \in S_{Y_c}} W_i} \leq \max\left\{\frac{k_1}{k_2}, \frac{k_2}{k_1}\right\}.$$

A.5 The Proof of Lemma 2.4.1

Proof:

$$E[h(x_1, \mu)] = \int \delta(x_1 - X_{i1}) E\left[\psi(Y_i - s)\frac{T_i}{\hat{\pi}(X_i)} \mid X_{i1}\right] g(X_{i1})\, dX_{i1}.$$

Since

$$\frac{1}{\hat{\pi}(X_i)} = \frac{1}{\pi(X_i)} + \frac{-1}{\hat{\pi}(X_i)\pi(X_i)}(\hat{\pi}(X_i) - \pi(X_i)).$$

We can rewrite $E[h(x_1, s)] = term1 - term2$.

Where

$$term1 = \int \delta_n(x_1 - X_{i1}) E\left[\psi(Y_i - s)\frac{T_i}{\pi(X_i)} \mid X_{i1}\right] g(X_{i1})\, dX_{i1},$$

and

$$term2 = \int \delta_n(x_1 - X_{i1}) E\left[\psi(Y_i - s)\frac{T_i}{\hat{\pi}(X_i)\pi(X_i)}(\hat{\pi}(X_i) - \pi(X_i)) \mid X_i\right] g(X_{i1})\, dX_{i1}.$$

Since in term2, $\psi(Y_i - s), T_i, \frac{1}{\hat{\pi}(X_i)\pi(X_i)}, \int \delta_n(x_1 - X_{i1}) dX_{i1}$ are all bounded, and by assumption (A3), we have $sup_{x \in \mathcal{X}} \mid \hat{\pi}(X) - \pi(X) \mid = O_p(\sqrt{n})$, we have $term2 \xrightarrow{P} 0$,
For term1, by Härdle (1984) Lemma 2.2, we have

$$\text{term1} \to E\left[\psi(Y_i)\frac{T_i}{\pi(X_i)} \mid X_{i1} = x_1\right] g(x_1).$$

Hence $E[h(x_1, s)] \xrightarrow{n \to \infty} H(x_1, s)$.
Let $H_{in}(X_i, s) = \psi(Y_i - s)\frac{T_i}{\hat{\pi}(X_i)}\delta_n(x_1 - X_{i1})$,then

$$Var(h(x_1, s)) = \frac{Var(H_{in}(X_i, s))}{n} = \frac{E(H_{in}(X_i, s)^2)}{n} - \frac{E(H_{in}(X_i, s))^2}{n},$$

where

$$
\begin{aligned}
E[H_{in}(x_1,s)^2] &= E\left[E\left[H_{in}(X_{\mathrm{i}},s)^2 \mid X_{\mathrm{i}1}\right]\right] \\
&= E\left[\left(\psi(Y_{\mathrm{i}}-s)\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}\delta_n(x_1-X_{\mathrm{i}1})\right)^2 \mid X_{\mathrm{i}1}\right] \\
&= \delta_n(x_1-X_{\mathrm{i}1})^2 E\left[\left(\psi(Y_{\mathrm{i}}-s)\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}\right)^2 \mid X_{\mathrm{i}1}\right].
\end{aligned}
$$

Again, we can expand the square of the inverse propensity score as:

$$
\frac{1}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})^2} = \frac{1}{\boldsymbol{\pi}(X_{\mathrm{i}})^2} + \frac{-2}{\xi^3(X_{\mathrm{i}})}(\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})-\boldsymbol{\pi}(X_{\mathrm{i}})),
$$

where $\xi(X_{\mathrm{i}})$ is a value between $\boldsymbol{\pi}(X_{\mathrm{i}}), \hat{\boldsymbol{\pi}}(X_{\mathrm{i}})$.

Then $E[H_{in}(x_1,s)^2]/n = term3 + term4$,

where

$$
\text{term3} = \frac{1}{n}\int \delta_n(x_1-X_{\mathrm{i}1})^2 E\left[\psi^2(Y_{\mathrm{i}}-s)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})^2} \mid X_{\mathrm{i}1}=x\right] g(x_1)\, dx,
$$

$$
\text{term4} = \frac{1}{n}\int \delta_n(x_1-X_{\mathrm{i}1})^2 E\left[\psi^2(Y_{\mathrm{i}}-s)T_{\mathrm{i}}\left(-\frac{2}{\xi(X_{\mathrm{i}})^3}\right)(\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})-\boldsymbol{\pi}(X_{\mathrm{i}})) \mid X_{\mathrm{i}1}=x\right] g(x_1)\, dx.
$$

Similar to term 1 and term 2, for term 4, we have ψ^2, T_{i} bounded, and hence we just need to show that

$(\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})-\boldsymbol{\pi}(X_{\mathrm{i}}))\delta_n(x_1-X_{\mathrm{i}1})^2/n \leq \mid \hat{\boldsymbol{\pi}}(X_{\mathrm{i}})-E[\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})] \mid \delta_n(x_1-X_{\mathrm{i}1})^2/n + \mid E[\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})]-\boldsymbol{\pi}(X_{\mathrm{i}}) \mid \delta_n(x_1-X_{\mathrm{i}1})^2/n \xrightarrow{P} 0.$

By the definition of δ_n, we have $\delta_n^2 \leq Ch^{-2}$.

For the first part, by chebyshev's inequality

$$
P\left(\mid \hat{\boldsymbol{\pi}}(X_{\mathrm{i}})-E[\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})] \mid \frac{C}{nh^2} \geq \epsilon\right) = P\left(\mid \hat{\boldsymbol{\pi}}(X_{\mathrm{i}})-E[\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})] \mid \geq \frac{\epsilon h^2 n}{C}\right) \leq \frac{\sigma_n^2 C^2}{\epsilon^2 h^2 n} \to 0.
$$

Under assumption $nh^2 \to \infty$ and $\sigma_n^2 = O(1/n)$, we have $\frac{\sigma_n^2 C^2}{\epsilon^2 h^2 n} \to 0$.

We have $\mid \hat{\boldsymbol{\pi}}(X_{\mathrm{i}})-E[\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})] \mid \frac{C}{nh^2} \xrightarrow{P} 0$.

Hence the first part converge to 0 in probability.

And for the second part, we assumed $\mid E[\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})] - \boldsymbol{\pi}(X_{\mathrm{i}}) \mid = O(1/n)$, we have $\mid E[\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})] - \boldsymbol{\pi}(X_{\mathrm{i}}) \mid \frac{C}{nh^2} \to 0$.

Hence, we have $term4 \xrightarrow{P} 0$.

For term3, define $\delta^*(u) = \frac{\delta_n^2}{\alpha_n(2)}$ where $\alpha_n(p) = \int \mid \delta_n(u) \mid^p du$, by Lemma 2.1 of (Hardle 1984) $\delta^*(u)$ is again a DFS.

Then

$$\text{term3} = \alpha_n(2) \int \delta_n^*(x_1 - X_{\mathrm{i}1}) E\left[\psi^2(Y_{\mathrm{i}} - s) \frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})^2} \mid X_{\mathrm{i}1} = x\right] g(x_1)\, dx.$$

Then

$$\frac{\text{term3}}{\alpha_n(2)} \xrightarrow{P} E\left[\psi^2(Y_{\mathrm{i}} - s) \frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})^2} \mid X_{\mathrm{i}1} = x\right] g(x_1),$$

and hence

$$\begin{aligned}\frac{\operatorname{Var}(H_{n\mathrm{i}}(x,s))}{\alpha_n(2)} \to & E\left[\psi^2(Y_{\mathrm{i}} - s) \frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})^2} \mid X_{\mathrm{i}1} = x_1\right] g(x_1) \\ & - \frac{1}{\alpha_\infty(2)} \left\{E\left[\psi(Y_{\mathrm{i}} - s) \frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} \mid X_{\mathrm{i}1} = x_1\right] g(x_1)\right\}^2,\end{aligned}$$

where $\alpha_\infty(2) = lim_{n\to\infty}\alpha_n(2)$ and by (Hardle 1984 Lemma2.1) we have $\alpha_\infty(2) = \infty$.

Hence

$$\frac{n \operatorname{Var}(H_n(x,s))}{\alpha_n(2)} \to E\left[\psi^2(Y_{\mathrm{i}} - s) \frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})^2} \mid X_{\mathrm{i}1} = x\right] g(x_1).$$

Then by Chebyshev Inequality

$$P(\mid h(x_1,s) - E(h(x_1,s)) \mid > \epsilon) \le \frac{Var(h(x_1,s))}{\epsilon^2}.$$

Assume $lim_{n\to\infty}\frac{\alpha_n(2)}{n} = 0$.

Then $\frac{nVar(H_n(x,s))}{\alpha_n(2)} \to C(x) < \infty$,

$Var(H_n(x,s)) \to 0$.

Then we have

$$lim_{n\to\infty} P(\mid h(x_1,s) - E(h(x_1,s)) \mid > \epsilon) = 0,$$

Hence $h(x_1, s) \xrightarrow{P} E[h(x_1, s)]$, and since $E[h(x_1, s)] \to H(x_1, s)$, we have $h(x_1, s) \xrightarrow{P} H(x_1, s)$.

A.6 The Proof of Theorem 2.4.1

Based on the continuous and strictly increased assumption of $\psi(\cdot)$, we have $H(x_1, s)$ and $h(x_1, s)$ also monotone and continuous.

By intermediate Theorem, for any $\epsilon > 0$ we have the event

$$h(x_1, c-\epsilon) < 0 \cap h(x_1, c+\epsilon) > 0 \text{ implies } c-\epsilon < \hat{\mu}_1(x_1) < c+\epsilon.$$

Then

$$P(h(x_1, c+\epsilon) < 0 \cap h(x_1, c-\epsilon) > 0) \leq P(c-\epsilon < \hat{\mu}_1(x_1) < c+\epsilon) = P(\mid \hat{\mu}_1(x_1) - c \mid < \epsilon).$$

Since we've showed $h(x_1, s) \xrightarrow{P} H(x_1, s)$,

we have

$$P(h(x_1, c-\epsilon) < 0 \cap h(x_1, c+\epsilon) > 0) \to P(H(x_1, c-\epsilon) < 0 \cap H(x_1, c+\epsilon) > 0) = 1.$$

Hence $lim_{n\to\infty} P(\mid \hat{\mu}_1(x_1) - c \mid < \epsilon) = 1 \Rightarrow \hat{\mu}_1(x_1) \xrightarrow{P} c$.

In Theorem 2.2.1, we've shown that $\mu_1(x_1)$ is a solution of $H(x_1, s)$, and there's only one solution for monotone continuous function, we have $\mu_1(x_1) = c$.

A.7 The Proof of Theorem 2.4.3

Proof: By the mean value Theorem, we have $\hat{\mu}_1(x_1) - \mu(x_1) = h(x_1, \mu_1(x_1))/D_n(x_1)$.

Where

$$D_n(x_1) = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} \delta_n(x_1 - X_{\mathrm{i}1}) \frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})} \psi'\left(Y_{\mathrm{i}} - \mu_1(x_1) - W_{\mathrm{i}}\left(\hat{\mu}_1(x_1) - \mu_1(x_1)\right)\right), \quad W_{\mathrm{i}} \in (0,1).$$

For the denominator $D_n(x)$, we can expand $D_n(x) = term5 + term6$.

Where

$term5 = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} \delta_n(x_1 - X_{\mathrm{i}1})\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'(Y_{\mathrm{i}} - \mu_1(x_1) - W_{\mathrm{i}}(\hat{\mu}_1(x_1) - \mu_1(x_1)))$,

$term6 = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} \delta_n(x_1 - X_{\mathrm{i}1})\frac{-T_{\mathrm{i}}}{\xi^2(X_{\mathrm{i}})}(\hat{\boldsymbol{\pi}}(X_{\mathrm{i}}) - \boldsymbol{\pi}(X_{\mathrm{i}}))\psi'(Y_{\mathrm{i}} - \mu_1(x_1) - W_{\mathrm{i}}(\hat{\mu}_1(x_1) - \mu_1(x_1)))$,

with $\epsilon(X_{\mathrm{i}}) \in (\boldsymbol{\pi}(X_{\mathrm{i}}), \hat{\boldsymbol{\pi}}(X_{\mathrm{i}}))$.

For term 6, similar to lemma 4.1, we have $term6 \xrightarrow{P} 0$.

For term 5, by $\hat{\mu}_1(x_1) \xrightarrow{P} \mu_1(x_1)$ and by WLLN, we have

$$E[\text{term5}] = E\left[\delta_n(x_1 - X_{\mathrm{i}1})\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'\left(Y_{\mathrm{i}} - \mu_1(x_1)\right)\right](1 + o(1)).$$

Then similar to what we did when proving consistency, by Härdle (1984) Lemma2.2, we have

$$\text{term5} \xrightarrow{P} E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'\left(Y_{\mathrm{i}} - \mu_1(x_1)\right) \mid X_{\mathrm{i}1} = x_1\right] g(x_1).$$

Then

$$D_n(x_1) \xrightarrow{P} E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'\left(Y_{\mathrm{i}} - \mu_1(x_1)\right) \mid X_{\mathrm{i}1} = x_1\right] g(x_1) = C(x_1)g(x_1).$$

For the numerator, from our proof of consistency, we have

$$h(x_1, \mu) = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}\left[\delta_n(X_{\mathrm{i}1} - x_1)\psi(Y_{\mathrm{i}} - s)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\right](1 + o_p(1)),$$

and

$$\frac{n}{\alpha_n(2)}\mathrm{Var}(h(x_1, \mu)) \to E\left[\psi(Y_{\mathrm{i}} - s)^2\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})^2} \mid X_{\mathrm{i}1} = x_1\right] g(x_1) = \sigma^2(x_1)g(x_1).$$

Hence, let $h^*(x_1) = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} \delta_n(X_{\mathrm{i}1} - x_1)\psi(Y_{\mathrm{i}} - s)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}$,

$H^*_{\mathrm{in}}(x_1) = \delta_n(X_{\mathrm{i}1} - x_1)\psi(Y_{\mathrm{i}} - s)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}$,

$B_n(x_1) = E(h^*(x_1))$,

$W_n(x_1) = \frac{(H^*_n(x_1) - B_n(x_1))}{(\frac{\alpha_n(2)}{n}\sigma^2(x_1)g(x_1))^{1/2}}$.

To verify the condition of Lyapunov CLT, we let $s_n^2 = \sum_{\mathrm{i}=1}^{n} Var(h^*(x_1))_{\mathrm{i}}^2$.

And for some $\eta > 0$, we have

$$\begin{aligned}
&\frac{1}{s_n^{2+\eta}} \sum_{\mathrm{i}=1}^{n} E\left[\mid H_{in}^*(x_1) - B_n(x_1) \mid^{2+\eta} \right] \\
= \;& \frac{n}{s_n^{2+\eta}} E\left[\mid H_{in}^*(x_1) - B_n(x_1) \mid^{2+\eta} \right] \\
= \;& \gamma_n \frac{n}{s_n^{2+\eta}} E\left[\delta^*(X_{\mathrm{i}1} - x_1) E\left[\mid \psi(Y_{\mathrm{i}} - s) \frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} - B_n(x_1) \mid^{2+\eta} \mid X_{\mathrm{i}1} \right] \right].
\end{aligned}$$

Where $\delta_n^*(\cdot) = \mid \delta_n(\cdot) \mid^{2+\eta} / \gamma_n$.

Similar to how we did in The Proof of lemma 2.4.1, under assumption (2) of this Theorem:

$$\frac{1}{s_n^{2+\eta}} \sum_{\mathrm{i}=1}^{n} E\left[\mid H_{in}^*(x_1) - B_n(x_1) \mid^{2+\eta} \right] = O\left(\frac{\gamma_n n}{s_n^{2+\eta}} \right) \to 0.$$

Then by Lyapunov CLT, we have $Z_n \xrightarrow{D} N(0,1)$.

A.8 The Proof of the Normality of $\hat{\tau}(x_1)$

For the normality, by our proof of the asymptotic properties of $h(x_1)$ we have

$$\begin{aligned}
\hat{\tau}(x_1) - \tau(x_1) &= (\hat{\mu}_1(x_1) - \mu_1(x_1)) - (\hat{\mu}_0(x_1) - \mu_0(x_1)) \\
&= \left(\frac{H_n^{*1}(x_1)}{D^1(x_1) g(x_1)} - \frac{H_n^{*0}(x_1)}{D^0(x_1) g(x_1)} \right) (1 + o_p(1)).
\end{aligned}$$

Where

$$H_n^{*1} = \frac{1}{nh} \sum_{\mathrm{i}=1}^{n} K\left(\frac{X_{\mathrm{i}1} - x_1}{h} \right) \left(\psi(Y_{\mathrm{i}} - \mu_1(x_1)) \frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} \right),$$

$$H_n^{*0} = \frac{1}{nh} \sum_{\mathrm{i}=1}^{n} K\left(\frac{X_{\mathrm{i}1} - x_1}{h} \right) \left(\psi(Y_{\mathrm{i}} - \mu_0(x_1)) \frac{1 - T_{\mathrm{i}}}{1 - \boldsymbol{\pi}(X_{\mathrm{i}})} \right).$$

$$D^1(x_1) = E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} \psi'(Y_{\mathrm{i}} - \mu_1(x_1)) \mid X_{\mathrm{i}1} = x_1 \right]$$

,

$$D^0(x_1) = E\left[\frac{1 - T_{\mathrm{i}}}{1 - \boldsymbol{\pi}(X_{\mathrm{i}})} \psi'(Y_{\mathrm{i}} - \mu_0(x_1)) \mid X_{\mathrm{i}1} = x_1 \right],$$

Similar to The Proof of the variance of $h(x_1)$ we have the variance

$$\frac{n}{\alpha_n(2)}\mathrm{Var}(\hat{\tau}(x_1)) \xrightarrow{n\to\infty} E\left[\left(\frac{H_n^{*1}(x_1)}{D^1(x_1)} - \frac{H_n^{*0}(x_1)}{D^0(x_1)}\right)^2 \mid X_{\mathrm{i}1} = x_1\right] g(x_1)^{-1}.$$

Where

$$H^{*1} = \psi(Y_{\mathrm{i}} - \mu_1(x_1))\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})},$$

$$H^{*0} = \psi(Y_{\mathrm{i}} - \mu_0(x_1))\frac{1 - T_{\mathrm{i}}}{1 - \boldsymbol{\pi}(X_{\mathrm{i}})}.$$

Verify the condition similarly to Theorem 4.2, by Lyapunov's CLT, we have the normality:

$$\sqrt{\frac{ng(x_1)}{\alpha_n(2)\sigma^2(x_1)}}(\hat{\tau}(x_1) - \tau(x_1)) \xrightarrow{D} N(0,1),$$

where

$$\sigma^2(x_1) = E\left[\left(\frac{H_n^{*1}(x_1)}{D^1(x_1)} - \frac{H_n^{*0}(x_1)}{D^0(x_1)}\right)^2 \mid X_{\mathrm{i}1} = x_1\right].$$

B. APPENDIX OF OUTLIER RESISTANT INFERENCE FOR HETEROGENEOUS TREATMENT EFFECT IN THE ABSENCE OF SYMMETRY AND LIGHT TAIL ASSUMPTIONS

For simplicity, we use μ_1^* to denote the true conditional mean function $\mu_1(x_1) = E[Y^{(1)} \mid X_1]$ and $\hat{\mu_1}$ to denote the estimated conditional mean function $\hat{\mu}_1(x_1)$.

B.1 The Proof of Theorem 3.3.1

Proof:

For purposes of exposition, we present the proof when X_1 is one-dimensional and with kernels of order 2. That is $l = 1$ and $s = 2$. All arguments can be generalized to higher dimensions or order cases.

Define our estimator $\hat{\mu}_1$ as the solution of estimating equation

$$l_{\hat{\boldsymbol{\pi}}}(\mu) = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} \psi_\alpha(Y_{\mathrm{i}} - \mu)\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})}K_h(X_{\mathrm{i}1} - x_1) = 0.$$

By the mean value theorem

$$\hat{\mu}_1 - \mu_1^* = \frac{l_{\hat{\boldsymbol{\pi}}}(\hat{\mu}_1) - l_{\hat{\boldsymbol{\pi}}}(\mu_1^*)}{l'_{\hat{\boldsymbol{\pi}}}(\tilde{\mu}_1)} = -\frac{l_{\hat{\boldsymbol{\pi}}}(\mu_1^*)}{l'_{\hat{\boldsymbol{\pi}}}(\tilde{\mu}_1)} = -\frac{l(\mu_1^*) + R_{\boldsymbol{\pi}}}{l'(\tilde{\mu}_1) + R'_{\boldsymbol{\pi}}}.$$

Where $\tilde{\mu}_1$ takes its value between $\hat{\mu}_1$ and μ_1^* and $R_{\boldsymbol{\pi}}, R'_{\boldsymbol{\pi}}$ are estimation errors of propensity sores with convergence rate $O_p(1/\sqrt{n})$.

Then to bound $\mid \hat{\mu}_1 - \mu_1^* \mid$, we only need to bound $\mid l_{\hat{\boldsymbol{\pi}}}(\mu_1^*) \mid \leq \mid l(\mu_1^*) \mid + \mid R_{\boldsymbol{\pi}} \mid$ from above and bound $\mid l'_{\hat{\boldsymbol{\pi}}}(\tilde{\mu}_1) \mid \geq \mid l'(\mu_1^*) \mid - \mid R'_{\boldsymbol{\pi}} \mid$ from below.

Begin with $\mid l(\mu_1^*) \mid$, we can bound it by bounding $\mid l(\mu_1^*) - E[l(\mu_1^*)] \mid$ and $E[l(\mu_1^*)]$ in the same time.

$$\begin{aligned} E[l(\mu_1^*)] &= E\left[\psi_\alpha(Y_{\mathrm{i}} - \mu_1^*)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}K_h(X_{\mathrm{i}1} - x_1)\right] \\ &= E\left[\psi_\alpha(Y^{(1)} - \mu_1^*)K_h(X_{\mathrm{i}1} - x_1)\right] \\ &= E\left[\psi_\alpha(Y^{(1)} - \mu_1^*) \mid X_1 = x_1\right] g(x_1) + O(h^2). \end{aligned}$$

Denote $v^{(1)} = Y^{(1)} - \mu_1^*$, then we have for any $k > 0$, $\delta > 0$

$$\begin{aligned} & \mid E\left[\frac{\psi_\alpha(v^{(1)})^k}{\boldsymbol{\pi}(X)^{k-1}} \mid X_1 = x_1\right] \mid \\ = & \mid E\left[\frac{v^{(1)k}1(\mid v^{(1)} \mid \leq \alpha)}{\boldsymbol{\pi}(X)^{k-1}} \mid X_1 = x_1\right] \mid + \mid E\left[\frac{\alpha^k 1(\mid v^{(1)} \mid > \alpha)}{\boldsymbol{\pi}(X)^{k-1}} \mid X_1 = x_1\right] \mid \\ = & \mid \alpha^{k-1-\delta}E\left[\frac{v^{(1)k}}{\alpha^{1-\delta}}\frac{1(\mid v^{(1)} \mid \leq \alpha)}{\boldsymbol{\pi}(X)^{k-1}} \mid X_1 = x_1\right] \mid + \mid E\left[\alpha^{1+\delta}\frac{1(\mid v^{(1)} \mid > \alpha)}{\boldsymbol{\pi}(X)^{k-1}} \mid X_1 = x_1\right] \mid \\ \leq & \alpha^{k-1-\delta}E\left[\mid v^{(1)} \mid^{1+\delta} \frac{1(\mid v^{(1)} \mid \leq \alpha)}{\boldsymbol{\pi}(X)^{k-1}} \mid X_1 = x_1\right] + E\left[\mid v^{(1)} \mid^{1+\delta} \frac{1(\mid v^{(1)} \mid > \alpha)}{\boldsymbol{\pi}(X)^{k-1}} \mid X_1 = x_1\right] \\ = & \alpha^{k-1-\delta}E\left[\mid v^{(1)} \mid^{1+\delta} / \boldsymbol{\pi}(X)^{k-1} \mid X_1 = x_1\right]. \end{aligned}$$

Then for $k = 1$

$$E[\psi_\alpha(v^{(1)}) \mid X_1] \leq E[\mid v^{(1)} \mid^{1+\delta} \mid X_1]\alpha^{-\delta},$$

which directly implies

$$\mid E(l(\mu_1^*)) \mid \leq E[\mid v^{(1)} \mid^{1+\delta} \mid X_1]g(x_1)\alpha^{1-\delta} + O(h^2).$$

And for $k = 2$

$$E[\psi_\alpha(v^{(1)})^2 \mid X_1] \leq E[\mid v^{(1)} \mid^{1+\delta} / \boldsymbol{\pi}(X)^{k-1} \mid X_1]\alpha^{1-\delta},$$

we'll use it in the following steps.

Next, to bound $\mid l(\mu_1^*) - El(\mu_1^*) \mid$, denote $\psi_{\alpha \mathrm{i}} = \psi_\alpha(v_{\mathrm{i}}^{(1)}) W_{\mathrm{i}}$ where $W_{\mathrm{i}} = \frac{T_{\mathrm{i}}}{\pi(X_{\mathrm{i}})} K_h(X_{\mathrm{i}1} - x_1)$, then by Bernstein's inequality, we have

$$P[l(\mu_1^*) - E(l(\mu_1^*)) \geq t] = P\left[\sum_{\mathrm{i}=1}^{n} (\psi_{\alpha \mathrm{i}} - E(\psi_{\alpha \mathrm{i}})) \geq nt\right] \leq exp\left(\frac{-(nt)^2}{2(n\sigma^2 + Mtn/3)}\right).$$

Where $M = sup \mid \psi_{\alpha \mathrm{i}} - E(\psi_{\alpha \mathrm{i}}) \mid \leq 2\alpha sup(W_{\mathrm{i}})$ and

$$\sigma^2 = \frac{1}{n} \sum_{\mathrm{i}=1}^{n} Var[\psi_{\alpha \mathrm{i}}] = E[\psi_{\alpha \mathrm{i}}^2] - E[\psi_{\alpha \mathrm{i}}]^2 \leq E[\psi_{\alpha \mathrm{i}}^2] \leq C_{K,n}(2) E[\frac{\mid v^{(1)} \mid^{1+\delta}}{\pi(X)}] \alpha^{1-\delta} g(x_1) + O(h)$$

let $\gamma = \frac{(nt)^2}{2(n\sigma^2 + Mtn/3)}$ and solve t from this expression

$$0 = t^2 - \frac{2M\gamma}{3n} t - \frac{2\gamma\sigma^2}{n}$$

we have

$$\begin{aligned} t &= \frac{M\gamma}{3n} \pm \sqrt{\frac{M^2\gamma^2}{9n^2} + \frac{\gamma\sigma^2}{n}} \\ &\leq \frac{\gamma}{3n} \alpha sup W_{\mathrm{i}} + \sqrt{\frac{(2\alpha sup(W_{\mathrm{i}}))^2 \gamma^2}{9n^2} + \frac{\gamma C_{K,n}(2) E[\frac{|v^{(1)}|^{1+\delta}}{\pi(X)}] \alpha^{1-\delta} g(x_1) + O(h^2)}{n}} := \delta_1. \end{aligned}$$

Then we have

$$P\left[\mid l(\mu_1^*) - E(\mu_1^*) \mid \geq \delta_1\right] \leq 2exp(-\gamma)$$

And for

$$\mid l(\mu_1^*) \mid \leq \mid E[l(\mu_1^*)] \mid + \mid l(\mu_1^*) - E[l(\mu_1^*)] \mid \leq E[\mid v^{(1)} \mid^{1+\delta} \mid X_1] \alpha^{-\delta} + \mid l(\mu_1^*) - E[l(\mu_1^*)] \mid$$

we have

$$P\left[\mid l(\mu_1^*) \mid \geq E[\mid v^{(1)} \mid^{1+\delta} \mid X_1] \alpha^{-\delta} + \delta_1\right] \leq 2exp(-\gamma).$$

Now we consider the error from the estimation of propensity score $R_\pi = O_p(1\sqrt{n})$. By the definition of $O_p(1/\sqrt{n})$, we have for any $\epsilon > 0$, there exists a finite constant $C_\pi > 0$ and a finite $N > 0$ such that for any $n > N_\pi$

$$P(\mid R_\pi \mid \geq C_\pi \frac{1}{\sqrt{n}}) \leq \epsilon.$$

We can use the following result to merge concentration inequalities. For two arbitrary continuous random variables A, B, we have $P(A \geq a) \leq x_a, P(B \geq b) \leq x_b$ we have

$$\begin{aligned} P(A + B \leq a + b) &\geq P(A \leq a \cap B \leq b) \\ &\geq P(A \leq a)P(B \leq b) \\ &= (1 - P(A \geq a))(1 - P(B \geq b)) \\ &\geq 1 - P(A \geq a) - P(B \geq b). \end{aligned}$$

And consequently, we have

$$P(A + B \geq a + b) \leq P(A \geq a) + P(B \geq b).$$

Take $\epsilon = exp(-\gamma)$, we have for $n > N_\pi$

$$P\left[\mid l(\mu_1^*) \mid + \mid R_\pi \mid \geq E[\mid v^{(1)} \mid^{1+\delta} \mid X_1]\alpha^{-\delta} + C_h h^2 + \delta_1 + C_\pi \frac{1}{\sqrt{n}}\right] \leq 3exp(-\gamma).$$

Where C_h is from the Taylor expansion of the mean of kernel estimator.

Next, we bound $\mid l'(\tilde{\mu}_1) \mid - \mid R'_\pi \mid$ from below, to do that, we need to bound $\mid l'(\tilde{\mu}_1) - E(l'(\tilde{\mu}_1)) \mid$ and $E[l'(\tilde{\mu}_1)]$ then include $\mid R'_\pi \mid$ in the inequality.

To bound $\mid E(l'(\tilde{\mu}_1)) \mid$ from below,

Denote $\tilde{v}_{\mathrm{i}}^{(1)} = Y_{\mathrm{i}}^{(1)} - \tilde{\mu}_1$, by Holder's inequality, we have

$$
\begin{aligned}
E\left[1\left(\mid \tilde{v}_{\mathrm{i}}^{(1)} \mid > \alpha\right)\frac{1}{h}K\left(\frac{X_{\mathrm{i}1}-x_1}{h}\right)\right] &= \frac{1}{\alpha}E\left[\alpha 1\left(\mid \tilde{v}_{\mathrm{i}}^{(1)} \mid > \alpha\right)\frac{1}{h}K\left(\frac{X_{\mathrm{i}1}-x_1}{h}\right)\right] \\
&\leq \frac{1}{\alpha}E\left[\mid \tilde{v}_{\mathrm{i}}^{(1)} \mid \frac{1}{h}K\left(\frac{X_{\mathrm{i}1}-x_1}{h}\right)\right] \\
&\leq \frac{1}{\alpha}E[1^2]^{1/2}\sqrt{E\left[\tilde{v}_{\mathrm{i}}^{(1)2}\frac{1}{h}K\left(\frac{X_{\mathrm{i}1}-x_1}{h}\right)\right]} \\
&= \frac{1}{\alpha}E\left[\tilde{v}_{\mathrm{i}}^{(1)2}\frac{1}{h}K\left(\frac{X_{\mathrm{i}1}-x_1}{h}\right)\right]^{1/2} \\
&\leq \frac{1}{\alpha}\left(C_{K,n}E\left[\tilde{v}_{\mathrm{i}}^{(1)2} \mid X_1 = x_1\right]g(x_1)\right)^{1/2} + O(h^{1/2}).
\end{aligned}
$$

Here we denote $\tilde{\sigma} = C_{K,n}^{1/2}E[\tilde{v}_{\mathrm{i}}^{(1)2} \mid X_1]^{1/2}$, then we have

$$
\frac{1}{g(x_1)} \mid E[\frac{1}{h}K(\frac{X_{\mathrm{i}1}-x_1}{h})] - E[l'(\tilde{\mu}_1)] \mid \leq \frac{\tilde{\sigma}}{\alpha} + O(\frac{\sqrt{h}}{\alpha}).
$$

And hence we have

$$
E[\mid l'(\tilde{\mu}_1) \mid] \geq (1 - \frac{\tilde{\sigma}}{\alpha} + O(\frac{\sqrt{h}}{\alpha}))g(x_1).
$$

Then to bound $\mid l'(\tilde{\mu}_1) - E[l'(\tilde{\mu}_1)] \mid$, denote $\psi'_{\alpha\mathrm{i}} = \psi'_\alpha(v_{\mathrm{i}}^{(1)})W_{\mathrm{i}}$.

By Bernstein's inequality, denote $Y_{\mathrm{i}} - \tilde{\mu}_1$ by $\tilde{v}_{\mathrm{i}}$ we have

$$
\begin{aligned}
& P\left[\mid l'(\tilde{\mu}_1) - E(l'(\tilde{\mu}_1)) \mid \leq t\right] \\
= \; & P\left[\mid \sum_{\mathrm{i}=1}^{n} 1(\mid \tilde{v}_{\mathrm{i}} \mid \leq \alpha)W_{\mathrm{i}} \mid \leq nt\right] \geq 1 - 2exp\left(-\frac{(nt)^2}{2(n\sigma_\alpha'^2 + M'_\alpha tn/3)}\right).
\end{aligned}
$$

Where

$$
M'_\alpha = sup \mid \psi'_{\alpha\mathrm{i}} \mid = sup \mid 1(\mid \tilde{v}_{\mathrm{i}} \leq \alpha \mid)W_{\mathrm{i}} \mid \geq sup(W_{\mathrm{i}}),
$$

and

$$
\sigma_\alpha'^2 = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}\mathrm{Var}\left[1\left(\mid \tilde{v}_{\mathrm{i}} \mid \leq \alpha\right)W_{\mathrm{i}}\right] \leq \frac{1}{n}\sum_{\mathrm{i}=1}^{n}\mathrm{Var}\left[W_{\mathrm{i}}\right] \leq C_{K,n}.
$$

Solve $\gamma = \frac{(nt)^2}{2(n\sigma_\alpha'^2 + M_\alpha' tn/3)}$ for t and plug in the solution to the inequality, we have

$$P\left[\mid l'(\tilde{\mu}_1) - E(l'(\tilde{\mu}_1)) \mid \leq \frac{\gamma}{3n} sup(W_{\mathrm{i}}) + \sqrt{\frac{M_\alpha'^2\gamma^2}{9n^2} + \frac{x\sigma_\alpha'^2}{n}}\right] \geq 1 - 2\mathrm{e}xp\left(-\gamma\right).$$

Let

$$\begin{aligned}\delta_2 := \quad & \frac{\gamma}{3n}\sup(W_{\mathrm{i}}) \\ + \quad & \sqrt{\frac{2\sup(W_{\mathrm{i}})^2\gamma^2}{9n^2} + \frac{\gamma}{n}\left(C_{K,n}(2)\frac{E\left[\frac{1}{\pi(X)} \mid X_1 = x_1\right]^{1/2}}{\alpha} g(x_1) + C_2'\frac{\sqrt{h}}{\alpha}\right)},\end{aligned}$$

we have $\frac{\gamma}{3n} sup(W_{\mathrm{i}}) + \sqrt{\frac{M'^2\gamma^2}{9n^2} + \frac{\gamma\sigma_\alpha'^2}{n}} \leq \delta_2$.

And since

$$\mid l'(\tilde{\mu}_1) - E(l'(\tilde{\mu}_1)) \mid = \mid E(l'(\tilde{\mu}_1)) - l'(\tilde{\mu}_1) \mid \geq \mid E(l'(\tilde{\mu}_1)) \mid - \mid l'(\tilde{\mu}_1) \mid,$$

we have

$$P\left[\mid l'(\tilde{\mu}_1) \mid \geq \mid E(l'(\tilde{\mu}_1)) \mid - \delta_2\right] \geq 1 - 2\mathrm{e}xp\left(-\gamma\right).$$

Together with the lower bound of $\mid E(l'(\tilde{\mu}_1)) \mid$, we have

$$P\left[\mid l'(\tilde{\mu}_1) \mid \geq (1 - \frac{\tilde{\sigma}}{\alpha})g(x_1) + C_1'\sqrt{h}/\alpha - \delta_2\right] \geq 1 - 2\mathrm{e}xp\left(-\gamma\right).$$

Assume

$$\begin{aligned}\max \quad & \left\{\frac{\tilde{\sigma}}{\alpha}g(x_1), 2C_1'\frac{\sqrt{h}}{\alpha}, \frac{\gamma}{3n}\sup(W_{\mathrm{i}}), \frac{2\sup(W_{\mathrm{i}})^2\gamma^2}{9n^2},\right. \\ & \left.\frac{\gamma}{n}C_{K,n}(2)\frac{E\left[\frac{1}{\pi(X)} \mid X_1 = x_1\right]^{1/2}}{\alpha} g(x_1)\right\} \\ \leq \quad & \frac{1}{8}g(x_1),\end{aligned}$$

there exist $C \geq \frac{1}{8}g(x_1)$ such that

$$P\left[\mid l'(\tilde{\mu}_1) \mid \geq C\right] \geq 1 - 2\mathrm{e}xp\left(-\gamma\right).$$

Treat R'_{π} similar to R_{π}, we have

$$P\left[\mid l'(\tilde{\mu}_1) \mid -R'_{\pi} \geq C - C'_{\pi}\frac{1}{\sqrt{n}}\right] \geq 1 - 3\mathrm{e}xp\left(-x\right).$$

Put the bound for the denominator and numerator together, we have

$$\begin{aligned} P[\mid \hat{\mu}_1 - \mu_1^* \mid \ \geq \ & \left(E[\mid v^{(1)} \mid^{1+\delta} \mid X_1]\alpha^{-\delta} + O(h^2) + \frac{2\gamma}{3n}\alpha \sup(W_{\mathrm{i}})\right) \\ & + \left(\sqrt{\frac{(2\alpha \sup(W_{\mathrm{i}}))^2\gamma^2}{9n^2} + \frac{\gamma C_{K,n}(2)E\left[\frac{|v^{(1)}|^{1+\delta}}{\pi(X)} \mid X_1 = x_1\right]\alpha^{1-\delta}g(x_1) + O(h^2)}{n}}\right) \\ & /C \\ < \ & 6\exp(-\gamma). \end{aligned}$$

And we can consider the conditions to minimize the upper bound. First, notice that the term $E[\mid v^{(1)} \mid^2 \mid X_1]\alpha^{-\delta}$ include $\alpha^{-\delta}$ while the terms $\frac{2x}{3n}\alpha sup(W_{\mathrm{i}})$ include α, and the term $\frac{(2\alpha sup(W_{\mathrm{i}}))^2\gamma^2}{9n^2}$ under the squared root sign is the same as $\frac{2x}{3n}\alpha sup(W_{\mathrm{i}})$ if it can dominate the other term under squared root sign. A direct way to find the optimal α is by calculating

$$E[\mid v^{(1)} \mid^{1+\delta} \mid X_1]\alpha^{-\delta} = \frac{2x}{3n}\alpha sup(W_{\mathrm{i}}),$$

and we get $\alpha = C(\frac{n}{\gamma})^{1/(\delta+1)}$ for a constant C.

However, with this rate, we have $E[\mid v^{(1)} \mid^2 \mid X_1]\alpha^{-\delta}$ and $\frac{2x}{3n}\alpha sup(W_{\mathrm{i}})$ equals $O((\frac{\gamma}{n})^{\delta/(\delta+1)})$ which dominated by

$$\sqrt{\frac{xC_{K,n}(2)(E[\frac{|v^{(1)}|^{1+\delta}}{\pi(X)}]\alpha^{1-\delta}g(x_1))}{n}} = O\left(\frac{1}{\sqrt{h}}\left(\frac{\gamma}{n}\right)^{\delta/(\delta+1)}\right)$$

Then to determine the growth rate of α that does not increasing the overall growth rate, we need,

$$E[\mid v^{(1)} \mid^{1+\delta} \mid X_1]\alpha^{-\delta} \quad = \quad O\left(\sqrt{\frac{xC_{K,n}(2)\left(E\left[\frac{|v^{(1)}|^{1+\delta}}{\pi(X)} \mid X_1\right]\alpha^{1-\delta}g(x_1)\right)}{n}}\right), \qquad \text{(B.1)}$$

and

$$\frac{2\gamma}{3n}\alpha \sup(W_{\mathrm{i}}) \quad = \quad O\left(\sqrt{\frac{xC_{K,n}(2)\left(E\left[\frac{|v^{(1)}|^{1+\delta}}{\pi(X)} \mid X_1\right]\alpha^{1-\delta}g(x_1)\right)}{n}}\right).$$

From which we get $C_{\alpha 1}(\frac{nh}{\gamma})^{\frac{1}{1+\delta}} \leq \alpha \leq C_{\alpha 2}(\frac{n}{\gamma h})^{\frac{1}{1+\delta}}$ for some constant $C_{\alpha 1}$ and $C_{\alpha 2}$.

To reach the highest robustness while maintaining the convergence rate, we consider selecting α by solving (B.1), which implies

$$\alpha = \left(\frac{nE[\mid v^{(1)|^{1+\delta}} \mid X_1 = x_1]^2}{xC_{K,n}(2)E[\mid v^{(1)} \mid^{1+\delta} /\boldsymbol{\pi}(X) \mid X_1 = x_1]g(x_1)}\right)^{1/(1+\delta)} = C\left(\frac{nh}{\gamma}\right)^{1/(1+\delta)}.$$

Also the bandwidth is required by $h^2 = O((\frac{\gamma}{nh})^{\delta/(1+\delta)})$ which implies

$$h = O((\frac{\gamma}{n})^{(\delta/(1+\delta))/(2+\delta/(1+\delta))}).$$

And to guarantee the convergence, we need $\gamma = o(nh)$.

For the bandwidth h, we need h^2 dominated by the other terms, that is $h^2 = o(\alpha^{-\delta})$, $\frac{2\gamma}{3n}\alpha$, $o(\sqrt{\alpha^{1-\delta}})$,$o(\frac{\gamma}{h}h^2)$. When $\gamma = o(nh)$ and based on our assumption $\delta = 1$, we only need $h = o(\alpha^{-1/2})$ and the other ones are satisfied. And then we can simplify the inequality as:

$$P[\mid \hat{\mu}_1 - \mu_1^* \mid \geq C\sqrt{\frac{\gamma}{nh}}] \leq 6exp(-\gamma).$$

For some constant C. Then the first inequality in the theorem is proved.

Next, let's prove the second inequality

Define $B(\hat{\mu}_1) = \frac{l(\hat{\mu}_1)-l(\mu_1^*)}{g(x_1)} - (\hat{\mu}_1 - \mu_1^*)$, then to prove the second inequality, we only have to bound $\mid B(\hat{\mu}_1) \mid$, to do that, we need to bound $\mid B(\hat{\mu}_1) - E[B(\hat{\mu}_1)] \mid$ and $\mid E[B(\hat{\mu}_1)] \mid$ separately.

By the mean value theorem

$$\begin{aligned} B(\hat{\mu}_1) &= \frac{l(\hat{\mu}_1) - l(\mu_1^*)}{g(x_1)} - (\hat{\mu}_1 - \mu_1^*) \\ &= \left(\frac{l'(\check{\mu}_1)}{g(x_1)} - 1\right)(\hat{\mu}_1 - \mu_1^*). \end{aligned}$$

Where $\check{\mu}_1$ takes its value between $\hat{\mu}_1$ and μ_1^*. Then to bound $\mid E[B(\hat{\mu}_1)] \mid$, we only have to bound $\mid \frac{l'(\check{\mu}_1)}{g(x_1)} - 1 \mid$ and $\mid \hat{\mu}_1 - \mu_1^* \mid$.

Denote $\check{\sigma} = E[\check{v}^{(1)2} \mid X_1]^{1/2}$, by our previous result, we have

$$P\left[UB \geq l'(\mu) \geq LB\right] \geq 1 - 2exp(-\gamma)$$

Where

$$UB = (1 - \frac{\check{\sigma}_p}{\alpha} + O(h^2))g(x_1) + \frac{\gamma}{3n}sup(W_{\mathrm{i}}) + \sqrt{\frac{2x\check{\sigma}'_\alpha}{n} + O(\frac{\gamma^2}{n^2})},$$

and

$$LB = (1 - \frac{\check{\sigma}_p}{\alpha} + O(h^2))g(x_1) - \frac{\gamma}{3n}sup(W_{\mathrm{i}}) - \sqrt{\frac{2x\check{\sigma}'_\alpha}{n} + O(\frac{\gamma^2}{n^2})},$$

$$\check{M}'_\alpha = sup \mid 1(\mid \check{v}_{\mathrm{i}} \leq \alpha \mid)W_{\mathrm{i}} \mid \geq sup(W_{\mathrm{i}}),$$

and

$$\begin{aligned} \check{\sigma}'^2_\alpha &= \frac{1}{n}\sum_{\mathrm{i}=1}^{n} Var[1(\mid \tilde{v}_{\mathrm{i}} \leq \alpha \mid)W_{\mathrm{i}}] \leq E[(1(\mid \tilde{v}_{\mathrm{i}} \leq \alpha \mid)W_{\mathrm{i}})^2] \\ &= C_{K,n}(2)E[1(\mid \tilde{v}_{\mathrm{i}} \leq \alpha \mid)/\boldsymbol{\pi}(X) \mid X_1 = x_1]g(x_1) + O(h) \\ &\leq C_{K,n}(2)\frac{E[\tilde{v}^{(1)2}/\boldsymbol{\pi}(X) \mid X_1 = x_1]^{1/2}}{\alpha}g(x_1) + O(h). \end{aligned}$$

Which implies

$$P\left[\mid \frac{l'(\mu_1)}{g(x_1)} - 1 \mid \leq (\frac{\check{\sigma}_p}{\alpha} + O(h^2)) + \frac{\gamma\alpha sup(W_{\mathrm{i}})}{3ng(x_1)} + \sqrt{\frac{x\check{\sigma}'_\alpha}{n} + O(\frac{\gamma^2}{n^2})}/g(x_1)\right] \geq 1 - 3exp(-\gamma).$$

Then together with our bound for $\mid \hat{\mu}_1 - \mu_1^* \mid$, we have

$$P\left[\mid \frac{l(\hat{\mu}_1) - l(\mu_1^*)}{g(x_1)} - (\hat{\mu}_1 - \mu_1^*) \mid \leq Bound1 \cdot Bound2\right] \geq 1 - 9exp(-\gamma),$$

where

$$Bound1 = (\frac{\check{\sigma}_p}{\alpha} + O(h^2)) - \frac{2x}{3ng(x_1)}\alpha sup(W_{\mathrm{i}}) - \sqrt{\frac{2x\check{\sigma}'_\alpha}{n} + O(\frac{1}{n^2})}/g(x_1),$$

$$Bound2 = (v_\delta \alpha^{-\delta} + O(h^2) + \frac{2x}{3n}\alpha sup(W_{\mathrm{i}}) + \sqrt{\frac{x(v_\delta \alpha^{1-\delta})g(x_1) + O(h^2)}{n}})/C.$$

Based on the conditions we've used, we have $Bound1 = O(\sqrt{\frac{\gamma}{nh}})$ and $Bound2 = O(\sqrt{\frac{\gamma}{nh}})$.

Then we have

$$P\left[\mid \frac{l(\hat{\mu}_1) - l(\mu_1^*)}{g(x_1)} - (\hat{\mu}_1 - \mu_1^*) \mid \leq C\frac{\gamma}{nh}\right] \geq 1 - 9exp(-\gamma).$$

Where C related to $\check{\sigma}_p$,sup(W_{i}),$\check{\sigma}'$, $E[\mid v^{(1)} \mid^2 \mid X_1]$ and $E[\mid v^{(1)} \mid^2 / \boldsymbol{\pi}(X) \mid X_1]$.

B.2 Propersition 3.2.1

Regarding the truncated random variable $\psi_\alpha(\epsilon_{\mathrm{i}})$, the following result shows that the differences between the first two moments of $\psi_\alpha(\epsilon_{\mathrm{i}})$ and ϵ_{i} depend on both α and the moments of ϵ_{i}. The higher moment ϵ_{i} has, the faster these differences decay as a function of α. We summarize this observation in the following proposition. The following result demonstrates that for the truncated random variable $\psi_\alpha(\epsilon_{\mathrm{i}})$, the differences between its first two moments and those of ϵ_{i} are dependent on both the parameter α and the moments of ϵ_{i}. Notably, as the higher moments of ϵ_{i} increase, these differences decay more rapidly as a function of α. This observation is summarized in the following proposition.

Propersition B.2.1. *Assume* $E[v^{(1)2}] > 0$ *and* $E[\mid v^{(1)} \mid^\kappa] < \infty$ *for some* $\kappa > 2$, *then*

$$\mid El_\alpha(v) \mid \leq \min(\frac{E[\mid v^{(1)} \mid^2 \mid X_1]}{\alpha} g(x_1) + C_\alpha h^2, \alpha^{1-\kappa} E[\mid v^{(1)} \mid^\kappa \mid X_1] g(x_1) + G_\alpha h^2),$$

and

$$\mid E[l_\alpha(v)^2] - E[l_2(v)^2] \mid \leq 2C_{K,n}(2)E[\frac{\mid v^{(1)} \mid^\kappa}{\boldsymbol{\pi}(X)} \mid X_1]g(x_1)(\kappa-2)^{-1}\alpha^{2-\kappa} + (C_\alpha + C_2)h.$$

Proof: Let $v = Y - \mu_1^*, v^{(1)} = Y^{(1)} - \mu_1^*$,
$l_2 = v\frac{T}{\boldsymbol{\pi}(X)}K_h(X_1 - x_1), l_\alpha(v) = \psi_\alpha(v)\frac{T}{\boldsymbol{\pi}(X)}K_h(X_1 - x_1)$.
Then we have
$E[l_2(v)] = C_2h^2, E[l_\alpha(v)] = E[\psi_\alpha(v^{(1)}) \mid X_1 = x_1]g(x_1) + C_\alpha h^2$,
where $C_2h^2, C_\alpha h^2$ are remainders of Taylor expansions.
Since $E[V^{(1)} \mid X_1 = x_1] = 0$, we have

$$\begin{aligned} E[\psi_\alpha(v^{(1)}) \mid X_1 = x_1] &= -E[(v^{(1)} - \alpha)1(v^{(1)} > \alpha) \mid X_1 = x_1] \\ &\quad +E[(-v^{(1)} - \alpha)1(v^{(1)} < -\alpha) \mid X_1 = x_1]. \end{aligned}$$

Hence, for any $2 \leq \iota \leq \kappa$,

$$\begin{aligned} \mid E[\psi_\alpha(v^{(1)}) \mid X_1 = x_1] \mid &\leq E[(\mid v^{(1)} \mid -\alpha)1(\mid v^{(1)} \mid > \alpha) \mid X_1 = x_1] \\ &\leq \alpha^{1-\iota}E[\mid v^{(1)} \mid^\iota \mid X_1 = x_1]. \end{aligned}$$

Which proofs the first result:

$$\mid E[l_\alpha(v)] \mid \leq \alpha^{1-\iota}E[\mid v^{(1)} \mid^\iota \mid X_1 = x_1]g(x_1) + G_\alpha h^2.$$

For the second inequality,

$$\begin{aligned} & E[\psi_\alpha(v^{(1)})^2/\boldsymbol{\pi}(X) \mid X_1 = x_1] \\ = \; & E[v^{(1)2}/\boldsymbol{\pi}(X) \mid X_1 = x_1] - \{E[v^{(1)2}/\boldsymbol{\pi}(X)1(\mid v^{(1)} \mid > \alpha) \mid X_1 = x_1] \\ & -\alpha^2 E[\frac{1}{\boldsymbol{\pi}(X)}1(\mid v^{(1)} \mid > \alpha) \mid X_1 = x_1]\}. \end{aligned}$$

Write $\eta = \mid v^{(1)} \mid$,

$$\begin{aligned}
& E[v^{(1)2}/\boldsymbol{\pi}(X)1(\mid v^{(1)} \mid > \alpha) \mid X_1 = x_1] \\
= \;& 2E[\int_0^\infty \frac{1(\eta > y)}{\boldsymbol{\pi}(X)} 1(\eta > \alpha) y dy \mid X_1 = x_1] \\
= \;& 2E[\int_0^\alpha \frac{1(\eta > \alpha)}{\boldsymbol{\pi}(X)} y dy \mid X_1 = x_1] + 2E[\int_\alpha^\infty \frac{1(\eta > y)}{\boldsymbol{\pi}(X)} y dy \mid X_1 = x_1] \\
= \;& 2E[\frac{1(\eta > \alpha)}{\boldsymbol{\pi}(X)} \int_0^\alpha y dy \mid X_1 = x_1] + 2E[\frac{1}{\boldsymbol{\pi}(X)} \int_\alpha^\infty 1(\eta > y) y dy \mid X_1 = x_1] \\
= \;& 2E[\frac{1(\eta > \alpha)}{\boldsymbol{\pi}(X)} \int_0^\alpha y dy \mid X_1 = x_1] + 2E[\frac{1}{\boldsymbol{\pi}(X)} \int_\alpha^\infty E[1(\eta > y) \mid X] y dy \mid X_1 = x_1] \\
= \;& 2E[\frac{1(\eta > \alpha)}{\boldsymbol{\pi}(X)} \int_0^\alpha y dy \mid X_1 = x_1] + 2E[\frac{1}{\boldsymbol{\pi}(X)} \int_\alpha^\infty P[\eta > y \mid X] y dy \mid X_1 = x_1].
\end{aligned}$$

By Markov's inequality:

$$P[\eta > y \mid X] \leq \frac{E[\eta^\kappa \mid X]}{y^\kappa}.$$

Then

$$\begin{aligned}
2E[\frac{1}{\boldsymbol{\pi}(X)} \int_\alpha^\infty P[\eta > y \mid X] y dy \mid X_1 = x_1] &\leq 2E[\frac{1}{\boldsymbol{\pi}(X)} \int_\alpha^\infty y^{1-\kappa} E[\eta^\kappa \mid X] dy \mid X_1 = x_1] \\
&= 2E[\frac{\eta^\kappa}{\boldsymbol{\pi}(X)} \mid X_1 = x_1] \int_\alpha^\infty y^{1-\kappa} dy \\
&= 2E[\frac{\eta^\kappa}{\boldsymbol{\pi}(X)} \mid X_1 = x_1](\kappa - 2)^{-1} \alpha^{2-\kappa}.
\end{aligned}$$

Then we have

$$\begin{aligned}
E[\frac{\eta^2}{\boldsymbol{\pi}(X)} 1(\eta > \alpha) \mid X_1 = x_1] &\leq \alpha^2 E[\frac{1(\mid v^{(1)} \mid > \alpha)}{\boldsymbol{\pi}(X)} \mid X_1 = x_1] \\
+2E[\frac{\eta^\kappa}{\boldsymbol{\pi}(X)} \mid X_1 = x_1](\kappa - 2)^{-1} \alpha^{2-\kappa}.
\end{aligned}$$

Since

$$E[l_2(v)^2] = C_{K,n}(2) E[\frac{\eta^2}{\boldsymbol{\pi}(X)} \mid X_1 = x_1] g(x_1) + C_2 h$$

we have

$$
\begin{aligned}
& E[l_\alpha(\eta)^2] - E[l_2(\eta)^2] \\
= \; & C_{K,n}(2)(-E[\frac{\eta^2}{\pi(X)}1(\eta > \eta) \mid X_1 = x_1] + \alpha^2 E[\frac{1(\eta > \alpha)}{\pi(X)} \mid X_1 = x_1])g(x_1) - (C_\alpha + C_2)h \\
\geq \; & -C_{K,n}(2)(\alpha^2 E[\frac{1(\mid v^{(1)} \mid > \alpha)}{\pi(X)} \mid X_1 = x_1] + 2E[\frac{\eta^\kappa}{\pi(X)} \mid X_1 = x_1](\kappa - 2)^{-1}\alpha^{2-\kappa} \\
& -\alpha^2 E[\frac{1(\mid v^{(1)} \mid > \alpha)}{\pi(X)} \mid X_1 = x_1])g(x_1) - (C_\alpha + C_2)h \\
= \; & -C_{K,n}(2)(2E[\frac{\eta^\kappa}{\pi(X)} \mid X_1 = x_1](\kappa - 2)^{-1}\alpha^{2-\kappa} \mid X_1 = x_1])g(x_1) - (C_\alpha + C_2)h.
\end{aligned}
$$

Which implies

$$
\begin{aligned}
\mid E\left[l_\alpha(v)^2\right] - E\left[l_2(v)^2\right] \mid \; \leq \; & C_{K,n}(2)\left[g(x_1)\left(2E\left[\frac{\eta^\kappa}{\pi(X)} \mid X_1 = x_1\right]\frac{1}{(\kappa-2)}\alpha^{2-\alpha}\right)\right. \\
& \left. +(C_\alpha + C_2)h\right].
\end{aligned}
$$

B.3 The Proof of Theorem 3.3.2

Proof: Let $W_n = n^{-1/2}\sum_{\mathrm{i}=1}^n \psi_\alpha(v_\mathrm{i})\frac{T_\mathrm{i}}{\pi(X_\mathrm{i})}K_h(X_{\mathrm{i}1} - x_1)/g(x_1)$,

$W_{0n} = n^{-1/2}\sum_{\mathrm{i}=1}^n\{\psi_\alpha(v_\mathrm{i})\frac{T_\mathrm{i}}{\pi(X_\mathrm{i})}K_h(X_{\mathrm{i}1} - x_1) - m_\alpha\}/g(x_1)$,

where $m_\alpha = E[\psi_\alpha(v_\mathrm{i})\frac{T_\mathrm{i}}{\pi(X_\mathrm{i})}K_h(X_{\mathrm{i}1} - x_1)]/g(x_1)$.

Define $\sigma_\alpha^2 = Var(\psi_\alpha(v)\frac{T}{\pi(X)}K_h(X_1 - x_1))/g(x_1)^2$, $\sigma^2 = Var(v\frac{T}{\pi(X)}K_h(X_1 - x_1))/g(x_1)^2$.

Let $T = \sqrt{n}(\hat{\mu}_1 - \mu_1^*)/\sigma$, $T_0 = \sigma_\alpha^{-1}W_{0n}$. First, we prove that T and T_0 are sufficiently close with high probability.

Note that

$$\begin{aligned} T - T_0 &= \sqrt{n}(\hat{\mu}_1 - \mu_1^*)/\sigma - \frac{1}{\sqrt{n}}\sum_{\mathrm{i}=1}^{n}\{\psi_\alpha(v_\mathrm{i})\frac{T_\mathrm{i}}{\boldsymbol{\pi}(X_\mathrm{i})}K_h(X_{\mathrm{i}1} - x_1) - m_\alpha\}/\sigma_\alpha g(x_1) \\ &= \frac{\sigma}{\sigma_\alpha}(\sqrt{n}(\hat{\mu}_1 - \mu_1^*)\frac{\sigma_\alpha\sigma}{\sigma^3} - \frac{W_n}{\sigma} + \frac{\sqrt{n}}{\sigma}m_\alpha + \frac{\sqrt{n}(\hat{\mu}_1 - \mu_1^*)}{\sigma} - \frac{\sigma^2}{\sigma^3}\sqrt{n}(\hat{\mu}_1 - \mu_1^*)) \\ &\leq \frac{\sigma}{\sigma_\alpha}\{\mid \sigma^2 - \sigma_\alpha^2 \mid \frac{\sqrt{n}}{\sigma^3} \mid \hat{\mu}_1 - \mu_1^* \mid + \frac{1}{\sigma} \mid \sqrt{n}(\hat{\mu}_1 - \mu_1^*) - W_n \mid + \frac{\sqrt{n}}{\sigma} \mid m_\alpha \mid\}. \end{aligned}$$

Then by Theorem 3.3.1 and Prop B.2.1, we can bound each term of the above bound and together as

$$P(\mid T - T_0 \mid \geq \delta_n) \leq 15\mathrm{e}^{-\gamma},$$

where $\delta_n = O(\frac{1}{n\alpha})O(\sqrt{\frac{\gamma}{nh}})/\sigma_\alpha + O(\frac{\gamma}{nh}) + O(\frac{\gamma}{\sqrt{nh}})$.

Then apply the Berry-Esseen inequality to T_0 and using (A.13), we have

$$\sup_{x\in R} \mid P(T_0 \leq x) - \Phi(x) \mid \leq \frac{C\rho_\alpha}{\sigma_\alpha^3\sqrt{n}}.$$

Since $T - T_0 \leq \delta_n$ implies $T \leq T_0 + \delta_n$, we have $P(T \geq x) = P(T_0 \geq x - \delta_n)$. Which implies

$$\sup_{x\in R} \mid P(T_0 \leq x) - \Phi(x) \mid \leq \frac{C\rho_\alpha}{\sigma_\alpha^3\sqrt{n}}.$$

To derive a Berry-Esseen bound for T, by the inequality $P(Y \leq a) \leq P(X \leq a + \epsilon) + P(\mid Y - X \mid \geq \epsilon)$ used in the proof of convergence of random variables, we have

$$P(T_0 \leq x - \delta_n) - P(\mid T - T_0 \mid > \delta_n) \leq P(T \leq x) \leq P(T_0 \leq x + \delta_n) + P(\mid T - T_0 \mid > \delta_n).$$

Then we have

$$\mid P(T \leq x) - \Phi(x) \mid \leq \max\{M1, M2\},$$

where

$$\begin{aligned} M1 &= P(T_0 \le x-\delta_n) - \Phi(x-\delta_n) - P(\mid T-T_0 \mid > \delta_n) + \Phi(x-\delta_n) - \Phi(x) \\ &\le \mid P(T \le x-\delta_n) - \Phi(x-\delta_n) \mid + P(\mid T-T_0 \mid > \delta_n) + \mid \Phi(x-\delta_n) - \Phi(x) \mid, \end{aligned}$$

$$\begin{aligned} M2 &= P(T_0 \le x+\delta_n) - \Phi(x+\delta_n) + P(\mid T-T_0 \mid > \delta_n) + \Phi(x+\delta_n) - \Phi(x) \\ &\le \mid P(T \le x+\delta_n) - \Phi(x+\delta_n) \mid + P(\mid T-T_0 \mid > \delta_n) + \mid \Phi(x+\delta_n) - \Phi(x) \mid . \end{aligned}$$

By the Berry-Essen bound,

$$\sup_{x\in R} \mid P(T_0 \le x+\delta_n) - \Phi(x+\delta_n) \mid \le \frac{CE[\mid l_\alpha(v_{\rm i}) - El_\alpha(v_{\rm i}) \mid^3]}{\sigma_\alpha^3\sqrt{n}},$$

and

$$\sup_{x\in R} \mid P(T_0 \le x-\delta_n) - \Phi(x-\delta_n) \mid \le \frac{CE[\mid l_\alpha(v_{\rm i}) - El_\alpha(v_{\rm i}) \mid^3]}{\sigma_\alpha^3\sqrt{n}}.$$

Also

$$\Phi(x+\delta_n) - \Phi(x) \le \frac{1}{\sqrt{2\pi}}\delta_n,$$

$$\Phi(x-\delta_n) - \Phi(x) \le \frac{1}{\sqrt{2\pi}}\delta_n.$$

We have

$$\mid P(T \le x) - \Phi(x) \mid \le 15{\rm e}^{-\gamma} + \frac{CE[\mid l_\alpha(v_{\rm i}) - E(l_\alpha(v_{\rm i})) \mid^3]}{\sigma_\alpha^3\sqrt{n}} + \frac{1}{\sqrt{2\pi}}\delta_n.$$

For σ_α, by the second inequality in proposition A2,

$$\mid E[l_\alpha(v)^2] - E[l_2(v)^2] \mid \le C_{K,n}(2)g(x_1)^2 E[\frac{\mid v^{(1)} \mid^\kappa}{\pi(X)} \mid X_1 = x_1](\kappa-2)^{-1}\alpha^{2-\kappa} + (C_\alpha + C_2)h.$$

Since $E[l_\alpha(v)^2] \le E[l_2(v)^2]$, we have

$$E[l_\alpha(v)^2] - E[l_2(v)^2] \ge -C_{K,n}(2)g(x_1)^2 E[\frac{\mid v^{(1)} \mid^\kappa}{\pi(X)} \mid X_1 = x_1](\kappa-2)^{-1}\alpha^{2-\kappa} - (C_\alpha + C_2)h.$$

Then further by the first inequality in proposition A2, we have

$$\begin{aligned} & E[l_\alpha(v)^2] - E[l_\alpha(v)]^2 - Var(l_2(v)) \\ \geq\ & -C_{K,n}(2)g(x_1)^2 E[\frac{\mid v^{(1)} \mid^\kappa}{\pi(X)} \mid X_1 = x_1](\kappa-2)^{-1}\alpha^{2-\kappa} - (C_\alpha + C_2)h \\ & -(\frac{E[\mid v^{(1)} \mid^2 \mid X_1 = x_1]}{\alpha} g(x_1) + C_\alpha h^2)^2, \end{aligned}$$

which implies

$$\begin{aligned} Var[l_\alpha(v)] \ \geq\ & Var(l_2(v)) - C_{K,n}(2)g(x_1)^2 E[\frac{\mid v^{(1)} \mid^\kappa}{\pi(X)} \mid X_1 = x_1](\kappa-2)^{-1}\alpha^{2-\kappa} - (C_\alpha + C_2)h \\ & -(\frac{E[\mid v^{(1)} \mid^2 \mid X_1 = x_1]}{\alpha} g(x_1) + C_\alpha h^2)^2, \end{aligned}$$

and suppose

$$\alpha \geq 8(C_{K,n}(2)g(x_1)^2 E[\frac{\mid v^{(1)} \mid^\kappa}{\pi(X)} \mid X_1 = x_1](\kappa-2)^{-1})1/(\kappa-2)/Var(l_2(v))$$

$$h \leq \frac{Var(l_2(v))}{8(C_\alpha + C_2)}$$

$$\alpha \geq 4(E[\mid v^{(1)} \mid^2 \mid X_1 = x_1]g(x_1))^{-1}/Var(l_2(v))$$

$$h \leq (Var(l_2(v))/4C_\alpha)^{1/4},$$

we have

$$\begin{aligned} C_{K,n}(2)g(x_1)^2 E[\frac{\mid v^{(1)} \mid^\kappa}{\pi(X)} \mid X_1 = x_1](\kappa-2)^{-1}\alpha^{2-\kappa} \ & - \ (C_\alpha + C_2)h \\ & - \ (\frac{E[\mid v^{(1)} \mid^2 \mid X_1 = x_1]}{\alpha} g(x_1) + C_\alpha h^2)^2 \\ \leq \ & \frac{Var(l_2(v))}{2}, \end{aligned}$$

which implies

$$Var[l_\alpha(v)] \geq Var(l_2(v))/2,$$

and we can further bound

$$\mid P(T \leq x) - \Phi(x) \mid \leq C\mathrm{e}^{-\gamma} + \frac{CE[\mid l_\alpha(v_\mathrm{i}) - E(l_2(v_\mathrm{i})) \mid^3]}{\sigma^3\sqrt{n}} + C\frac{\gamma}{\sqrt{n}h} \leq C\{\frac{\gamma}{\sqrt{n}h} + \mathrm{e}^{-\gamma}\}.$$

To guarantee the convergence, we need $x \to \infty$ as $n \to \infty$ and $x = o(\sqrt{n}h)$. When $h = O(n^{1/5})$, we have $x = o(n^{3/10})$.

And similarly, for $T_1 = \sqrt{n}(\hat{\mu}_1 - \mu_1^*)/\sigma_\alpha$, with $T_1 - T_0 \leq \frac{1}{\sigma_\alpha} \mid \sqrt{n}(\hat{\mu}_1 - \mu_1^*) - W_n \mid + \frac{\sqrt{n}}{\sigma_\alpha} \mid m_\alpha \mid := \delta_{n2}$, we have

$$\mid P(T_1 \leq x) - \Phi(x) \mid \leq 9\mathrm{e}^{-\gamma} + \frac{CE[\mid l_\alpha(v_\mathrm{i}) - E(l_\alpha(v_\mathrm{i})) \mid^3]}{\sigma_\alpha^3\sqrt{n}} + \frac{1}{\sqrt{2\pi}}\delta_{n2} \leq C_2(\mathrm{e}^{-\gamma} + \frac{\gamma}{\sqrt{n}h}).$$

B.4 The Proof of Theorem 3.4.1

We've shown $\hat{\mu}_1$ and W_n are asymptotic normal, their counterpart in the control group can be similarly shown, hence the estimator of CATE $\hat{\tau}(x_1) = \hat{\mu}_1(x_1) - \hat{\mu}_0(x_1)$ is also asymptotic normal and intuitively we can approximate $\sqrt{n}(\hat{\tau}(x_1) - \tau(x_1))$ by

$$W_\tau = n^{-1/2}\sum_{\mathrm{i}=1}^{n}\{A_\mathrm{i}(x_1) - B_\mathrm{i}(x_1) - m_{\alpha_1} + m_{\alpha_0}\}$$

. **Proof:** Similar to the proof of Thm2, we can write $\mid P(T_\tau \leq x) - \Phi(x) \mid \leq max\{M_{1\alpha}, M_{2\alpha}\}$.

Where

$$\begin{aligned} M_{1\alpha} \leq \quad & \mid P(T_{joint\alpha} \leq x - \delta_{joint,\alpha,n}) - \Phi(x - \delta_{joint,\alpha,n}) \mid + P(\mid T_\tau - T_{joint\alpha} \mid > \delta_{joint,\alpha,n}) \\ & + \mid \Phi(x - \delta_{joint,\alpha,n}) - \Phi(x) \mid \end{aligned}$$

$$\begin{aligned} M_{2\alpha} \leq \quad & \mid P(T_{joint\alpha} \leq x + \delta_{joint,\alpha,n}) - \Phi(x + \delta_{joint,\alpha,n}) \mid + P(\mid T_\tau - T_{joint\alpha} \mid > \delta_{joint,\alpha,n}) \\ & + \mid \Phi(x + \delta_{joint,\alpha,n}) - \Phi(x) \mid . \end{aligned}$$

Here we focus on $M_{1\alpha}$ since $M_{2\alpha}$ can be treated similarly.

For $\mid P(T_{joint\alpha} \leq x - \delta_{\text{joint},\alpha,n}) - \Phi(x - \delta_{\text{joint},\alpha,n}) \mid$, by Berry-Esseen theorem, we have

$$\begin{aligned} sup_x \mid P(T_{joint\alpha} &\leq x - \delta_{\text{joint},\alpha,n}) - \Phi(x - \delta_{\text{joint},\alpha,n}) \mid \\ &\leq \frac{C\rho_{\text{joint},3}}{\sigma_{\tau,\alpha}(x_1)^3\sqrt{n}}. \end{aligned}$$

where $\rho_{joint,3} = E[\mid A_{\text{i}}(x_1) - B_{\text{i}}(x_1) - E[A_{\text{i}}(x_1) - B_{\text{i}}(x_1)] \mid^3]$.

For the second term $P(\mid T_\tau - T_{joint\alpha} \mid > \delta_{\text{joint},\alpha,n})$, by Prop A2, we have

$$\begin{aligned} &\mid T_\tau - T_{joint\alpha} \mid \\ \leq\ & \frac{1}{\sigma_{\tau,\alpha}(x_1)}(\mid \sqrt{n}(\hat{\mu}_1 - \mu_1^*) - W_{n1} \mid + \mid \sqrt{n}(\hat{\mu}_0 - \mu_0^*) - W_{n0} \mid) + \frac{\sqrt{n}}{\sigma_{\tau,\alpha}(x_1)}(\mid m_{\alpha_1} \mid + \mid m_{\alpha_0} \mid) \\ \leq\ & \delta_{\alpha,n}, \end{aligned}$$

where $\delta_{\alpha,n} = O(\frac{\gamma}{\sqrt{nh}})/\sigma_{\tau,\alpha}(x_1)$. Then we have

$$P(\mid T_\tau - T_{joint\alpha} \mid \geq \delta_{\text{joint}})/\sigma_{joint\alpha} \leq 18\text{e}^{-\gamma}.$$

For the third term, we have $\mid \Phi(x - \delta_{\text{joint},\alpha,n}) - \Phi(x) \mid \leq \frac{1}{\sqrt{2\pi}}\delta_{\text{joint}\alpha}/\sigma_{\tau,\alpha}(x_1)$.

Then we only have to replace $\sigma_{\tau,\alpha}(x_1)$ by a form without α_1 or α_0. Since

$$\begin{aligned} & Var(A_{\text{i}}(x_1) - B_{\text{i}}(x_1)) \\ =\ & Var(A_{\text{i}}(x_1)) + Var(B_{\text{i}}(x_1)) - 2Cov(A, B) \\ \geq\ & Var(A_{\text{i}}(x_1)) + Var(B_{\text{i}}(x_1)) - 2 \mid Cov(A, B) \mid \\ \geq\ & Var(A_{\text{i}}(x_1)) + Var(B_{\text{i}}(x_1)) - 2 \mid E(AB) \mid -2 \mid E(A_{\text{i}}(x_1)) \mid\mid E(B_{\text{i}}(x_1)) \mid . \end{aligned}$$

And since $T(1 - T) = 0$, we have

$$\begin{aligned} E(AB) &= E[\psi_{\alpha_1}(v_{\text{i}})\frac{T}{\boldsymbol{\pi}(X)}K_h(X_{\text{i}1} - x_1) \cdot \psi_{\alpha_1}(v_{\text{i}})\frac{1 - T}{1 - \boldsymbol{\pi}(X)}K_h(X_{\text{i}1} - x_1)] \\ &= E[\psi_{\alpha_1}(v_{\text{i}})\psi_{\alpha_1}(v_{\text{i}})\frac{T}{\boldsymbol{\pi}(X)}\frac{1 - T}{1 - \boldsymbol{\pi}(X)}K_h^2(X_{\text{i}1} - x_1)] = 0. \end{aligned}$$

and by prop 2, $\mid E(A_{\text{i}}(x_1)) \mid\mid E(B_{\text{i}}(x_1)) \mid \leq \frac{v_2^2}{\alpha_1^2\alpha_2^2}$,

we can bound the variance from below

$$\sigma_{\tau,\alpha}(x_1)^2 = Var(A_{\mathrm{i}}(x_1) - B_{\mathrm{i}}(x_1)) \geq \frac{\sigma_1^2}{2} + \frac{\sigma_2^2}{2} - \frac{v_2^2}{\alpha_1^2\alpha_2^2}.$$

By the condition we required for α in Thm1, $\alpha_1 \geq \sqrt{\frac{nh}{\gamma}}$ and $\alpha_0 \geq C\sqrt{\frac{nh}{\gamma}}$, we have

$$\sigma_{\tau,\alpha}(x_1)^2 = Var(A_{\mathrm{i}}(x_1) - B_{\mathrm{i}}(x_1)) \geq \frac{\sigma_1^2}{2} + \frac{\sigma_2^2}{2} - C\frac{v_2^2\gamma^2}{n^2h^2}.$$

Then we can have a similar result as Thm2 for $\hat{\tau}$:

$$sup_x \mid P(T_\tau \leq x) - \Phi(x) \mid \leq C(\frac{\gamma}{\sqrt{n}h} + \mathrm{e}^{-\gamma}),$$

and when $x = o(\sqrt{n}h)$ we have $T_\tau \xrightarrow{d} N(0,1)$.

C. APPENDIX OF NONPARAMETRIC OUTLIER RESISTANT CONDITIONAL AVERAGE TREATMENT EFFECT ESTIMATOR WITH SUFFICIENT DIMENSION REDUCTION

C.1 Technical conditions

For the proof of asymptotic theorems of kernel based M-estimator, Härdle (1984) use $\delta_n(\cdot)$ to denote the delta function sequence(DFS), satisfying

(D1) $\int \mid \delta_n(u) \mid du < \infty$, for all n,

(D2) $\int \delta_n(u)du = 1$,

(D3) $\delta_n(u) \to 0$, uniformly on $\mid u \mid > \eta, \eta > 0$ as $n \to \infty$,

(D4) $\int_{|u|>\eta} \delta_n(u)du \to 0$, for each $\eta > 0$ as $n \to \infty$.

Our choice of kernel $K(u) = \frac{1}{h_n}K(u/h_n)$ can be regarded as a DFS.

Their lemma 2.1 shows that $\{\delta_{n,p}(u)\} = \{\mid \delta_n(u) \mid^p /\alpha_n(p)\}$ is also a DFS, where $\alpha_n(p) = \int \mid \delta_n(u) \mid^p du < \infty$ with $\alpha_n(p) \to \infty$. And their lemma 2.2 shows that for an integrable and continuous function $h(u)$ we have $h(\cdot)\delta_n(\cdot)$ integrable and $\int h(u)\delta_n(u)du \to h(0)$ as $n \to \infty$. For convenience, we'll use these notation and lemmas in our proof.

The following regularity assumptions are used in the deriving of asymptotic properties.

(A3) Let $\psi(u)$ be monotone, anti-symmetric, bounded, and having two continuous bounded derivatives with $\psi'(0) > 0$.

(A4) Let $f(Y \mid X_1)$ the conditional probability density function of Y given X_1 be symmetric and have bounded partial derivative on $x_1 \in \chi$ a Cartesian product of compact intervals of the real line. The density of X_1, is assumed to be $\inf_{x\in\chi} g(x_1) \geq c_0 > 0$ with second derivative exists.

(A5) $\sup_{x\in\chi} E[\psi(Y^{(\mathrm{j})} - \mu_{\mathrm{j}})^2 \mid X = x] < \infty$ for $\mathrm{j} = 0, 1$. The functions $m_{\mathrm{j}}(x) = E[\psi(Y^{(\mathrm{j})} - \mu_{\mathrm{j}}) \mid X = x], \mathrm{j} = 0, 1$ and $f(x)$ are s-times continuously differentiable on χ.

(A6) The population propensity score $\boldsymbol{\pi}(X)$ is s times continuously differentiable on χ.

(A7) $K(u)$ is a kernel of order s, is symmetric around zero, is equal to 0 outside $\prod_{\mathrm{i}=1}^{k}[-1, 1]$, and is continuously differentiable. $K_1(u)$ is a kernel of order s_1 is symmetric around 0, and is s times continuously differentiable.

(A8) The bandwidths h, h_1 satisfy the following conditions as $n \to \infty$:

(i) $h \to 0$ and $\log(n)/(n^{k+s}) \to 0$.

(ii) $nh_1^{2s_1+l} \to 0$ and $nh_1^l \to \infty$.

(iii)$h^{2s}h_1^{-2s-l} \to 0$ and $nh_1^l h^{2s} \to 0$.

C.2 The Proof of Theorem 4.3.1

Lemma C.2.1. *Given assumption (A1) to (A8), we have* $H_n(x_1) \xrightarrow{P} H(x_1)$ *for each* $x_1 \in \chi$.

Proof: We can write it as

$$H_n(x_1, \mu_1(X_1)) = H1 + H2 + H3.$$

Where

$$H_1 = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n}\psi(Y_\mathrm{i} - \mu_1(x_1))\frac{T_\mathrm{i}}{\boldsymbol{\pi}(X_\mathrm{i})}K_1\left(\frac{X_{\mathrm{i}1} - x_1}{h_1}\right),$$

$$H_2 = \frac{-1}{nh_1}\sum_{\mathrm{i}=1}^{n}\psi(Y_\mathrm{i} - \mu_1(x_1))\frac{T_\mathrm{i}}{\boldsymbol{\pi}(X_\mathrm{i})^2}K_1\left(\frac{X_{\mathrm{i}1} - x_1}{h_1}\right)(\hat{\boldsymbol{\pi}}(X_\mathrm{i}) - \boldsymbol{\pi}(X_\mathrm{i})),$$

$$H_3 = \frac{2}{nh_1}\sum_{\mathrm{i}=1}^{n}\psi(Y_\mathrm{i} - \mu_1(x_1))\frac{T_\mathrm{i}}{\boldsymbol{\pi}^*(X_\mathrm{i})^3}K_1\left(\frac{X_{\mathrm{i}1} - x_1}{h_1}\right)(\hat{\boldsymbol{\pi}}(X_\mathrm{i}) - \boldsymbol{\pi}(X_\mathrm{i}))^2,$$

with $\boldsymbol{\pi}^*(X_\mathrm{i})$ between $\boldsymbol{\pi}(X_\mathrm{i})$ and $\hat{\boldsymbol{\pi}}(X_\mathrm{i})$.

Denote $\omega_{\mathrm{ij}} = \frac{\frac{1}{nh^k}K\left(\frac{X_\mathrm{i} - X_\mathrm{j}}{h}\right)}{\frac{1}{nh^k}\sum_{t:t\neq\mathrm{i}}K\left(\frac{X_\mathrm{i} - X_t}{h}\right)}$,

$\hat{\boldsymbol{\pi}}(X_\mathrm{i}) = \sum_{\mathrm{j}:\mathrm{j}\neq\mathrm{i}}\omega_{\mathrm{ij}}T_\mathrm{j}$,

$\Psi_{\boldsymbol{\pi}} = -\psi(Y_\mathrm{i} - \mu_1(x_1))\frac{T_\mathrm{i}}{\boldsymbol{\pi}(X_\mathrm{i})^2}$,

$S_{\boldsymbol{\pi}} = E[\Psi_{\boldsymbol{\pi}} \mid X_\mathrm{i}] = -E[\psi(Y_\mathrm{i}^{(1)} - \mu_1(x_1)) \mid X_\mathrm{i}]\frac{1}{\boldsymbol{\pi}(X_\mathrm{i})}$,

$\zeta = \Psi_{\boldsymbol{\pi}} - S_{\boldsymbol{\pi}}(X_\mathrm{i})$,

$\epsilon_\mathrm{i} = T_\mathrm{i} - \boldsymbol{\pi}(X_\mathrm{i})$,

$\beta_n = E[\hat{\boldsymbol{\pi}}(X_\mathrm{i}) \mid X_1, ..., X_n] - \boldsymbol{\pi}(X_\mathrm{i}) = \sum_{\mathrm{j}:\mathrm{j}\neq\mathrm{i}}\omega_{\mathrm{ij}}\boldsymbol{\pi}(X_\mathrm{j}) - \boldsymbol{\pi}(X_\mathrm{i})$.

Then

$$\begin{aligned} H2 &= \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right) S_\pi(X_\mathrm{i})(\hat{\boldsymbol{\pi}}(X_\mathrm{i})-\boldsymbol{\pi}(X_\mathrm{i})) \\ &\quad +\frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right) \zeta_\mathrm{i}(\hat{\boldsymbol{\pi}}(X_\mathrm{i})-\boldsymbol{\pi}(X_\mathrm{i})) \\ &= H21 + H22, \end{aligned}$$

where

$$\begin{aligned} H21 &= \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right) S_\pi(X_\mathrm{i})\left(\sum_{\mathrm{j}:\mathrm{j}\neq\mathrm{i}} \omega_{\mathrm{ij}} T_\mathrm{j} - \boldsymbol{\pi}(X_\mathrm{i})\right) \\ &= \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right) S_\pi(X_\mathrm{i})\left(\sum_{\mathrm{j}:\mathrm{j}\neq\mathrm{i}} \omega_{\mathrm{ij}}(\epsilon_\mathrm{j}+\boldsymbol{\pi}(X_\mathrm{j})) - \boldsymbol{\pi}(X_\mathrm{i})\right) \\ &= \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right) S_\pi(X_\mathrm{i})\epsilon_\mathrm{i} \\ &\quad +\frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right) S_\pi(X_\mathrm{i})\left(\sum_{\mathrm{j}:\mathrm{j}\neq\mathrm{i}} \omega_{\mathrm{ij}}\epsilon_\mathrm{j} - \epsilon_\mathrm{i}\right) \\ &\quad +\frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right) S_\pi(X_\mathrm{i})\left(\sum_{\mathrm{j}:\mathrm{j}\neq\mathrm{i}} \omega_{\mathrm{ij}}\boldsymbol{\pi}(X_\mathrm{j}) - \boldsymbol{\pi}(X_\mathrm{i})\right) \\ &= \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right) S_\pi(X_\mathrm{i})\epsilon_\mathrm{i} \\ &\quad +\frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} \epsilon_\mathrm{i}\left(\sum_{\mathrm{j}:\mathrm{j}\neq\mathrm{i}} K_1\left(\frac{X_{1\mathrm{j}}-x_1}{h_1}\right) S_\pi(X_\mathrm{j}) - K_1\left(\frac{X_{1\mathrm{j}}-x_1}{h_1}\right) S_\pi(X_\mathrm{i})\right) \\ &\quad +\frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right) S_\pi(X_\mathrm{i})\left(\sum_{\mathrm{j}:\mathrm{j}\neq\mathrm{i}} \omega_{\mathrm{ij}}\boldsymbol{\pi}(X_\mathrm{j}) - \boldsymbol{\pi}(X_\mathrm{i})\right) \\ &= H211 + H212 + H213, \end{aligned}$$

$$
\begin{aligned}
& H1 + H211 \\
= \; & \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n}\psi(Y_{\mathrm{i}}-\mu_1(x_1))\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)+\frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)S_{\boldsymbol{\pi}}(X_{\mathrm{i}})\epsilon_{\mathrm{i}} \\
= \; & \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\left(\psi(Y_{\mathrm{i}}-\mu_1(x_1))\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}-E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\frac{T_{\mathrm{i}}-\boldsymbol{\pi}(X_{\mathrm{i}})}{\boldsymbol{\pi}(X_{\mathrm{i}})}\right) \\
= \; & \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\left(\left(\psi(Y_{\mathrm{i}}-\mu_1(x_1))-E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\right)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\right. \\
& \left.+E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\right).
\end{aligned}
$$

The term $H3, H22, H212, H213$ can be bounded the same way as Abrevaya, Hsu, and Lieli (2015) with its lemma6.1 and assumption 5,6 and 7(ii),8(iii),8(i),8(ii).

Then we have

$$
\begin{aligned}
H_n =& \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\left(\left(\psi(Y_{\mathrm{i}}-\mu_1(x_1))-E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\right)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\right. \\
& \left.+E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\right)+o_p\left(\frac{1}{\sqrt{nh_1}}\right) \\
=& H_n^*+o_p\left(\frac{1}{\sqrt{nh_1}}\right),
\end{aligned}
$$

$$
\begin{aligned}
E[H_n] =& E\left[\frac{1}{h_1}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\left((\psi(Y_{\mathrm{i}}-\mu_1(x_1))-E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\right)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\right. \\
& \left.+\; E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\right)\Big]+o\left(\frac{1}{\sqrt{nh_1}}\right).
\end{aligned}
$$

By Härdle (1984) lemma 2.2, we have

$$
\begin{aligned}
E[H_n] \to & E\left[\left((\psi(Y_{\mathrm{i}}-\mu_1(x_1))-E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\right)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\right. \\
& \left.+\; E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\right)\mid X_1=x_1\Big]\,g(x_1) \\
& = E\left[\psi(Y_{\mathrm{i}}-\mu_1(x_1))\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\mid X_1=x_1\right]g(x_1).
\end{aligned}
$$

Let

$$H_{in} = \frac{1}{h_1} K_1\left(\frac{X_{i1} - x_1}{h_1}\right)\left(\left(\psi(Y_i - \mu_1(x_1)) - E[\psi(Y_i^{(1)} - \mu_1(x_1)) \mid X_i]\right)\frac{T_i}{\boldsymbol{\pi}(X_i)} + E[\psi(Y_i^{(1)} - \mu_1(x_1)) \mid X_i]\right),$$

then

$$\begin{aligned}
\mathrm{Var}(H_n) &= \mathrm{Var}(H_n^* + o_p\left(\frac{1}{\sqrt{nh_1}}\right)) \\
&= \mathrm{Var}(H_n^*) + \mathrm{Var}\left(o_p\left(\frac{1}{\sqrt{nh_1}}\right)\right) + \mathrm{Cov}\left(H_n^*, o_p\left(\frac{1}{\sqrt{nh_1}}\right)\right) \\
&= \mathrm{Var}(H_n^*) + E\left[o_p\left(\frac{1}{\sqrt{nh_1}}\right)^2\right] - E\left[o_p\left(\frac{1}{\sqrt{nh_1}}\right)\right]^2 \\
&\quad + E\left[H_n^* o_p\left(\frac{1}{\sqrt{nh_1}}\right)\right] - E[H_n^*]E\left[o_p\left(\frac{1}{\sqrt{nh_1}}\right)\right] \\
&= \mathrm{Var}(H_n^*) + o\left(\frac{1}{\sqrt{nh_1}}\right),
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(H_n^*) &= \frac{1}{n}\mathrm{Var}(H_{ni}^*) \\
&= \frac{1}{n}\left(E\left[H_{ni}^{*2}\right] - E\left[H_{ni}^*\right]^2\right),
\end{aligned}$$

$$\begin{aligned}
E[H_{ni}^{*2}] &= E\left[\frac{1}{h_1^2} K_1\left(\frac{X_{i1} - x_1}{h_1}\right)^2\left(\left(\psi(Y_i - \mu_1(x_1)) - E[\psi(Y_i^{(1)} - \mu_1(x_1)) \mid X_i]\right)\frac{T_i}{\boldsymbol{\pi}(X_i)}\right.\right. \\
&\quad \left.\left. + E[\psi(Y_i^{(1)} - \mu_1(x_1)) \mid X_i]\right)^2\right].
\end{aligned}$$

Denote

$$\sigma^2(x_1)g(x_1) = E\left[\left(\left(\psi(Y_i - \mu_1(x_1)) - E[\psi(Y_i^{(1)} - \mu_1(x_1)) \mid X_i]\right)\frac{T_i}{\boldsymbol{\pi}(X_i)} + E[\psi(Y_i^{(1)} - \mu_1(x_1)) \mid X_i]\right)^2 \mid X_{i1} = x_1\right] g(x_1).$$

Then by Härdle (1984) lemma 2.1 and 2.2, we have $\frac{1}{\alpha_n}E[H_{ni}^{*2}] \xrightarrow{P} \sigma^2(x_1)g(x_1)$,

and hence $\frac{n}{\alpha_n(2)}Var(H_n) \to \sigma^2(x_1)g(x_1)$.

Then by Chebyshev Inequality, we have $H_n \xrightarrow{P} E[H_n]$.

C.3 The Proof of Theorem 4.3.1

Proof:

By the mean value theorem, we have $\hat{\mu}_1(x_1) - \mu_1(x_1) = \frac{H_n(x,\mu_1(x_1))}{D_n(x_1)}$.

Where

$$H_n(x_1) = \frac{1}{nh_1}\sum_{i=1}^{n}\psi(Y_i - \mu_1(x_1))\frac{T_i}{\hat{\pi}(X_i)}K_1\left(\frac{X_{i1} - x_1}{h_1}\right),$$

and

$$D_n(x_1) = \frac{1}{nh_1}\sum_{i=1}^{n}\psi'\left(Y_i - \mu_1(x_1) - W_i(\hat{\mu}_1(x_1) - \mu_1(x_1))\right)\frac{T_i}{\hat{\pi}(X_i)}K_1\left(\frac{X_{i1} - x_1}{h_1}\right), W_i \in (0,1).$$

For the denominator $D_n(x_1)$, since

$$\frac{1}{\hat{\pi}(X_i)} = \frac{1}{\pi(X_i)} + \frac{1}{\pi(X_i)\hat{\pi}(X_i)}(\hat{\pi}(X_i) - \pi(X_i)).$$

We can expand $D_n(x_1) = D1 + D2$, where

$$D1 = \frac{1}{nh_1}\sum_{i=1}^{n}K_1\left(\frac{X_{i1} - x_1}{h_1}\right)\frac{T_i}{\pi(X_i)}\psi'\left(Y_i - \mu_1(x_1) - W_i(\hat{\mu}_1(x_1) - \mu_1(x_1))\right),$$

$$\begin{aligned} D2 =& \frac{1}{nh_1}\sum_{i=1}^{n}K_1\left(\frac{X_{i1} - x_1}{h_1}\right)\frac{-T_i}{\pi(X_i)\hat{\pi}(X_i)}(\hat{\pi}(X_i) - \pi(X_i)) \\ & \cdot \psi'\left(Y_i - \mu_1(x_1) - W_i(\hat{\mu}_1(x_1) - \mu_1(x_1))\right). \end{aligned}$$

For the term D1, we have

$$\begin{aligned} E[D1] &= E\left[\frac{1}{h_1}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)\right] \\ &= E\left[\frac{1}{h_1}E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)\mid X_{\mathrm{i}1}\right]K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\right] \\ &= \int E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)\mid X_{\mathrm{i}1}\right]g(X_{\mathrm{i}1})\frac{1}{h_1}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)dX_{\mathrm{i}1}. \end{aligned}$$

Follows the Lemma 2.2 in Härdle (1984), we have

$$E[D1]\to E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)\mid X_{\mathrm{i}1}=x_1\right]g(x_1)\mathrm{as}n\to\infty.$$

Similarly, $Var[D1]=\frac{1}{n}Var[D1_{\mathrm{i}}]=\frac{1}{n}(E[D1_{\mathrm{i}}^2]+E[D1_{\mathrm{i}}]^2)$.

And we have

$$\begin{aligned} \frac{E[D1_{\mathrm{i}}^2]}{n} &= \frac{1}{n}E\left[\frac{1}{h_1^2}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)^2\frac{T_{\mathrm{i}}^2}{\boldsymbol{\pi}(X_{\mathrm{i}})^2}\psi'\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)^2\right] \\ &= \frac{1}{n}\int E\left[\frac{T_{\mathrm{i}}^2}{\boldsymbol{\pi}(X_{\mathrm{i}})^2}\psi'\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)^2\mid X_{\mathrm{i}1}\right]g(X_{\mathrm{i}1})\frac{1}{h_1^2}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)^2dX_{\mathrm{i}1}. \end{aligned}$$

By Lemma 2.1 in Härdle (1984), we have that $\delta_2(u)=\left(\frac{1}{h_1}K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\right)^2/\alpha_n(2)$ is again a DFS. Furthermore, by Lemma 2.2 in Härdle (1984), we obtain

$$\frac{n}{\alpha_n(2)}E[D1_{\mathrm{i}}^2]\to E\left[\frac{T_{\mathrm{i}}^2}{\boldsymbol{\pi}(X_{\mathrm{i}})^2}\psi'\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)^2\mid X_{\mathrm{i}1}=x_1\right]g(x_1),$$

as $n\to\infty$.

Assume $\alpha_n(2)/n\to 0$ then we have $Var[D1]\to 0$.

Hence, by Chebyshev's inequality, we have $D_1\xrightarrow{P}E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)\mid X_{\mathrm{i}1}=x_1\right]g(x_1)$.

For the term $D2$, since $\hat{\boldsymbol{\pi}}(X_{\mathrm{i}})\xrightarrow{P}\boldsymbol{\pi}(X_{\mathrm{i}})$ and the remaining parts are bounded, it follows that $D2\xrightarrow{P}0$.

Thus, we have $D_n(x_1)\xrightarrow{P}E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)\mid X_{\mathrm{i}1}=x_1\right]g(x_1)$.

For the numerator, let

$$H_n^* = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\left[\left(\psi\left(Y_{\mathrm{i}}-\mu_1(x_1)\right)-E\left[\psi\left(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1)\right)\mid X_{\mathrm{i}}\right]\right)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} + E\left[\psi\left(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1)\right)\mid X_{\mathrm{i}}\right]\right].$$

Define $B_n(x_1) = E\left[H_n^*\right]$.

Then, let

$$W_n(x_1) = \frac{H_n^* - B_n(x_1)}{\left(\frac{\alpha_n(2)}{n}\sigma^2(x_1)g(x_1)^{-1}\right)^{1/2}}.$$

To show the normality of $\hat{\mu}_1(x_1) - \mu_1(x_1)$, it suffices to demonstrate that $W_n(x_1) \xrightarrow{P} N(0,1)$.

Since

$$\frac{n}{\alpha_n(2)}\mathrm{Var}(H_n(x_1) - E[H_n(x_1)]) = \frac{n}{\alpha_n(2)}\mathrm{Var}(H_n(x_1)) \to \sigma^2(x_1)g(x_1),$$

by applying Liapunov's Central Limit Theorem (CLT) and mimicking the proof of Theorem 3.5 in Pagan et al. (1999), we have

$$W_n(x_1) \xrightarrow{D} N(0,1).$$

C.4 Proof of Theorem 4.4.2

Proof:

$$\begin{aligned} E[Y^{(1)}T \mid B^T X] &= E[E[Y^{(1)}T \mid X] \mid B^T X] \\ &= E[E[Y^{(1)} \mid X]E[T \mid X] \mid B^T X]. \end{aligned}$$

Here if we have $E[Y^{(1)} \mid X] = E[Y^{(1)} \mid B^T X]$, then

$$\begin{aligned} E[E[Y^{(1)} \mid X]E[T \mid X] \mid B^T X] &= E[E[Y^{(1)} \mid B^T X]E[T \mid X] \mid B^T X] \\ &= E[Y^{(1)} \mid B^T X]E[E[T \mid X] \mid B^T X] \\ &= E[Y^{(1)} \mid B^T X]E[T \mid B^T X]. \end{aligned}$$

Similarly, we can show the dimension-reduced version of unconfoundedness assymption hold with $E[T \mid X] = E[T \mid B^T X]$.

C.5 Proof of Theorem 4.4.3

Proof:

When X is replaced by $S = \hat{B}^T X$ we have $\hat{\mu}_1(x_1) - \mu_1(x_1) = \frac{H_n(x_1)}{D_n(x_1)}$.

Where

$$H_n(x_1) = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} \psi(Y_\mathrm{i} - \mu_1(x_1))\frac{T_\mathrm{i}}{\hat{\boldsymbol{\pi}}(\hat{B}^T X_\mathrm{i})}K_1(\frac{X_{\mathrm{i}1} - x_1}{h_1}),$$

$$D_n(x_1) = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} \psi'(Y_\mathrm{i} - \mu_1(x_1) - W_\mathrm{i}(\hat{\mu}_1(x_1) - \mu_1(x_1)))\frac{T_\mathrm{i}}{\hat{\boldsymbol{\pi}}(\hat{B}^T X_\mathrm{i})}K_1(\frac{X_{\mathrm{i}1} - x_1}{h_1}), W_\mathrm{i} \in (0,1).$$

We write $H_n(x_1)$ and $D_n(x_1)$ them in short as

$$H_n(x_1) = E_n[\psi(Y_\mathrm{i} - \mu_1(x_1))\frac{T_\mathrm{i}}{\hat{\boldsymbol{\pi}}(\hat{B}^T X_\mathrm{i})} \mid X_1 = x_1]$$

and

$$D_n(x_1) = E_n[\psi'(Y_\mathrm{i} - \mu_1(x_1) - W_\mathrm{i}(\hat{\mu}_1(x_1) - \mu_1(x_1)))\frac{T_\mathrm{i}}{\hat{\boldsymbol{\pi}}(\hat{B}^T X_\mathrm{i})} \mid X_1 = x_1].$$

Follows W. Luo, Y. Zhu, and Ghosh (2017), we assume $E_n[\frac{\partial\psi(Y-\mu_1(x_1))T}{\partial\hat{\boldsymbol{\pi}}(B^T X)} \mid X_1 = x_1] = 0$.

This can be satisfied if we let $\alpha = \frac{E_n[X\frac{\partial\psi(Y-\mu_1(x_1))/\hat{\boldsymbol{\pi}}(B^T X)}{\partial B^T X}]}{E_n[\frac{\partial\psi(Y-\mu_1(x_1))/\hat{\boldsymbol{\pi}}(B^T X)}{\partial B^T X}]}$, and use $B^T(X - \alpha)$ in place of $B^T X$.

Then

$$
\begin{aligned}
H_n(x_1) &= E_n[\psi(Y_{\mathrm{i}} - \mu_1(x_1))\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(B^T X_{\mathrm{i}})} \mid X_1 = x_1] \\
&\quad + E_n[\frac{\partial \psi(Y - \mu_1(x_1))T}{\partial \hat{\boldsymbol{\pi}}(B^T X)} \mid X_1 = x_1](\hat{B} - B) + O_p(\mid\mid \hat{B} - B \mid\mid^2).
\end{aligned}
$$

Follows W. Luo, Y. Zhu, and Ghosh (2017), when the condition 7 in W. Luo, Y. Zhu, and Ghosh (2017) holds, we have $\mid \|\hat{B} - B\| \mid = o_p(n^{-1/4})$, hence

$$
\sqrt{nh_1} H_n(x_1) = E_n[\psi(Y_{\mathrm{i}} - \mu_1(x_1))\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(B^T X_{\mathrm{i}})} \mid X_1 = x_1] + o_p(1).
$$

And the denominator,

$$
\begin{aligned}
D_n(x_1) &= E_n[\psi'(Y_{\mathrm{i}} - \mu_1(x_1) - W_{\mathrm{i}}(\hat{\mu}_1(x_1) - \mu_1(x_1)))\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(B^T X_{\mathrm{i}})} \mid X_1 = x_1] + O_p(\mid\mid \hat{B} - B \mid\mid) \\
&= E_n[\psi'(Y_{\mathrm{i}} - \mu_1(x_1) - W_{\mathrm{i}}(\hat{\mu}_1(x_1) - \mu_1(x_1)))\frac{T_{\mathrm{i}}}{\hat{\boldsymbol{\pi}}(B^T X_{\mathrm{i}})} \mid X_1 = x_1] + o_p(1).
\end{aligned}
$$

Then similar to the case without SDR, we can show the consistency and normality with $Var(\mu_1(x_1)) = \frac{\alpha_n(2)}{ng(x_1)} \frac{E[((\psi(Y - \mu_1(x_1)) - E[\psi(Y^{(1)} - \mu_1(x_1))|B^T X])\frac{T}{\pi(B^T X)} + E[\psi(Y^{(1)} - \mu_1(x_1))|B^T X])^2 | X_1 = x_1]}{E[\frac{T}{\pi(B^T X)}\psi'(Y - \mu_1(x_1))|X_1 = x_1]^2}$.

C.6 Proof of the asymptotic distribution of $\hat{\tau}(x_1)$

Since $\hat{\mu}_1(x_1)$ and $\hat{\mu}_0(x_1)$ are consistent estimators of $\mu_1(x_1)$ and $\mu_0(x_1)$, respectively, we have:

$$
\hat{\tau}(x_1) = \hat{\mu}_1(x_1) - \hat{\mu}_0(x_1) \xrightarrow{P} \mu_1(x_1) - \mu_0(x_1) = \tau(x_1).
$$

Based on our proof of the asymptotic properties of $H_n(x_1)$, we have:

$$
\begin{aligned}
\hat{\tau}(x_1) - \tau(x_1) &= (\hat{\mu}_1(x_1) - \mu_1(x_1)) - (\hat{\mu}_0(x_1) - \mu_0(x_1)) \\
&= \frac{H_n^{*1}(x_1)}{D^1(x_1)g(x_1)} - \frac{H_n^{*0}(x_1)}{D^0(x_1)g(x_1)} + o_p\left(\frac{1}{\sqrt{nh_1}}\right),
\end{aligned}
$$

Where

$$H_n^{*1} = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\Big[\Big(\psi(Y_{\mathrm{i}}-\mu_1(x_1)) - E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\Big)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} + E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\Big],$$

$$H_n^{*0} = \frac{1}{nh_1}\sum_{\mathrm{i}=1}^{n} K_1\left(\frac{X_{\mathrm{i}1}-x_1}{h_1}\right)\Big[\Big(\psi(Y_{\mathrm{i}}-\mu_0(x_1)) - E[\psi(Y_{\mathrm{i}}^{(0)}-\mu_0(x_1))\mid X_{\mathrm{i}}]\Big)\frac{1-T_{\mathrm{i}}}{1-\boldsymbol{\pi}(X_{\mathrm{i}})} + E[\psi(Y_{\mathrm{i}}^{(0)}-\mu_0(x_1))\mid X_{\mathrm{i}}]\Big],$$

$$D^1(x_1)g(x_1) = E\left[\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'(Y_{\mathrm{i}}-\mu_1(x_1))\mid X_{\mathrm{i}1}=x_1\right]g(x_1),$$

$$D^0(x_1)g(x_1) = E\left[\frac{1-T_{\mathrm{i}}}{1-\boldsymbol{\pi}(X_{\mathrm{i}})}\psi'(Y_{\mathrm{i}}-\mu_0(x_1))\mid X_{\mathrm{i}1}=x_1\right]g(x_1).$$

Similar to the proof of the variance of $H_n(x_1)$ we have:

$$\frac{n}{\alpha_n(2)}\mathrm{Var}(\hat{\tau}(x_1)) \xrightarrow{P} E\left[\left(\frac{H^{*1}(x_1)}{D^1(x_1)} - \frac{H^{*0}(x_1)}{D^0(x_1)}\right)^2 \mid X_1 = x_1\right] g(x_1)^{-1},$$

where

$$H^{*1} = \Big(\psi(Y_{\mathrm{i}}-\mu_1(x_1)) - E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}]\Big)\frac{T_{\mathrm{i}}}{\boldsymbol{\pi}(X_{\mathrm{i}})} + E[\psi(Y_{\mathrm{i}}^{(1)}-\mu_1(x_1))\mid X_{\mathrm{i}}],$$

$$H^{*2} = \Big(\psi(Y_{\mathrm{i}}-\mu_0(x_1)) - E[\psi(Y_{\mathrm{i}}^{(0)}-\mu_0(x_1))\mid X_{\mathrm{i}}]\Big)\frac{1-T_{\mathrm{i}}}{1-\boldsymbol{\pi}(X_{\mathrm{i}})} + E[\psi(Y_{\mathrm{i}}^{(0)}-\mu_0(x_1))\mid X_{\mathrm{i}}].$$

Then by Lyapunov's CLT, we have the normality:

$$\sqrt{\frac{ng(x_1)}{\alpha_n(2)\sigma^2}}(\hat{\tau}(x_1)-\tau(x_1)) \xrightarrow{D} N(0,1),$$

where

$$\begin{aligned}
\sigma^2 =& E\left[\left(\frac{H^{*1}(x_1)}{D^1(x_1)} - \frac{H^{*0}(x_1)}{D^0(x_1)}\right)^2 \mid X_1 = x_1\right] \\
=& \frac{\alpha_n(2)}{ng(x_1)}\left(E\left[\frac{\mathrm{Var}(\psi(Y^{(1)} - \mu_1(x_1)))}{\boldsymbol{\pi}(X)D^1(x_1)} \mid X_1 = x_1\right]\right. \\
&+ \left. E\left[\frac{\mathrm{Var}(\psi(Y^{(0)} - \mu_0(x_1)))}{(1-\boldsymbol{\pi}(X))D^0(x_1)} \mid X_1 = x_1\right]\right) \\
&+ \mathrm{Var}\left(\frac{E[\psi(Y^{(1)} - \mu_1(x_1)) \mid X]}{D^1(x_1)} - \frac{E[\psi(Y^{(0)} - \mu_0(x_1)) \mid X]}{D^0(x_1)} \mid X_1 = x_1\right).
\end{aligned}$$

VITA

Ran Mo

Education Background

Ph.D. Candidate in Applied Statistics, Indiana University-Purdue University Indianapolis

Expected Graduation: December 2024

M.S. in Applied Statistics, Indiana University-Purdue University Indianapolis

August 2018 - May 2019

B.S. in Information and Computing Science, Beijing University of Technology

September 2009 - July 2013

Teaching Experience

Math 11000: Fundamentals of Algebra (Spring 2021 & Summer 2021)

Math 11100: Intermediate Algebra (Autumn 2021)

STAT 30100: Elementary Statistical Methods I (Spring 2022)

Graded and tutored for various courses.

Conducted problem-solving sessions in probability and statistics.

Awards

Department of Mathematical Sciences Advanced Statistics Graduate Student Award, 2024

Conferences

Presented at the ENAR 2023 Spring Meeting.

Poster presentation at the 2023 Midwest Biopharmaceutical Statistics Workshop (MBSW).